



# 2025 年（第 18 届）中国大学生计算机 设计大赛

## 大数据实践赛作品报告

作品编号： 2025043615

作品名称： 一盾当关：基于多模态的诈骗风险监测系统

填写日期： 2025.04.01

### 填写说明：

- 1、正文、标题格式已经在本文中设定，请勿修改；标题的快捷键为“Ctrl+#”，正文快捷键为“Ctrl+0”；
- 2、本支撑文档应结构清晰，突出重点，适当配合图表，描述准确，不易冗长拖沓；
- 3、提交支撑时，以 PDF 格式提交；
- 4、本支撑内容是正式参赛内容的组成部分，务必真实填写。如不属实，将导致奖项等级降低甚至取消本作品参加比赛；
- 5、如果使用大模型，请参赛者在提交的作品中详细说明大模型在研究或创作中的具体使用方式以及优化的开发模块。

# 目录

第 1 章 作品概述 .....	1
第 2 章 问题分析 .....	2
2.1 问题来源 .....	2
2.1.1 新型引导式诈骗领域现状 .....	2
2.1.2 诈骗的技术方法 .....	3
2.1.3 诈骗防御挑战 .....	6
2.2 现有解决方案 .....	7
2.2.1 针对诱导式诈骗的识别技术 .....	7
2.2.2 针对伪造内容的诈骗识别技术 .....	10
2.3 本作品要解决的痛点问题 .....	12
2.3.1 防范多样化诈骗手段的性能差 .....	12
2.3.2 抵御实时交互式诈骗的技术少 .....	12
2.3.3 判断诈骗行为的解释信服度低 .....	12
2.3.4 多角度检测系统的可嵌入性弱 .....	12
2.4 解决的思路 .....	13
2.4.1 作品功能与性能需求 .....	13
2.4.2 数据集 .....	14
第 3 章 技术方案 .....	26
3.1 诈骗风险内容识别系统设计方案 .....	26
3.1.1 设计思路 .....	26
3.1.2 技术路线总览 .....	26
3.1.3 系统功能设计架构 .....	28
3.1.4 系统设计 .....	30

---

3.2 关键技术与原理.....	35
3.2.1 基于跨模态时序不一致性的伪造检测算法 .....	35
3.2.2 基于交叉模态情绪一致性的诈骗心理识别方法 .....	42
3.2.3 基于文本诱导性特征捕捉的可解释性检测技术 .....	55
3.2.4 基于大数据技术的用户风险形象分析.....	64
 第 4 章 系统实现.....	66
4.1 前端架构 .....	66
4.2 系统部署 .....	69
4.3 模块训练 .....	71
4.3.1 数据来源 .....	71
4.3.2 数据训练 .....	73
4.3.3 改进过程 .....	74
4.4 系统展示 .....	75
4.4.1 实时通话检测.....	75
4.4.2 离线文件检测.....	77
4.4.3 风险智能分析.....	78
4.4.4 用户风险形象分析 .....	81
4.4.5 信息交流论坛.....	82
 第 5 章 测试分析.....	83
5.1 系统性能测试.....	83
5.1.1 通话内容 .....	83
5.1.2 通话行为 .....	86
5.1.3 通话载体 .....	97
5.2 系统功能测试 .....	108
5.3 系统易用测试 .....	110
5.4 系统可靠测试 .....	111

---

第6章 作品总结.....	113
6.1 作品特色与创新点.....	113
6.1.1 推出全面且高效的多模态风险内容识别系统 .....	113
6.1.2 基于跨模态时序不一致性的人脸伪造检测算法 .....	113
6.1.3 基于交叉模态情绪一致性的诈骗心理识别方法 .....	113
6.1.4 高度可解释的诈骗风险检测解释方案.....	114
6.1.5 面向风险形象的大数据推送.....	114
6.1.6 面向应用的易部署性、易拓展性、易维护性 .....	114
6.2 应用推广 .....	115
6.2.1 社交网络 .....	115
6.2.2 司法取证 .....	116
6.2.3 金融领域 .....	116
6.2.4 电商与市场营销 .....	117
6.3 作品展望 .....	117
参考文献 .....	118

---

## 第1章 作品概述

随着深度伪造技术与生成式人工智能的爆发式发展，诈骗手段已从传统低效的、广撒网的话术诱导演变为新型高效的、个性化的“技术伪造 + 心理操控”深度融合的引导式诈骗。诈骗分子会通过伪造熟人音视频、融合社会工程学技术，在实时交互中构建“量身定制”的高质量骗局，其隐蔽性和成功率大幅提升。同样，随着技术发展，目前也涌现出各种有效的检测技术，例如针对音视频伪造的检测、异常动作捕捉、情绪分析等，这些技术的完善研发，能为诈骗防御带来新的指导方向和可能手段。然而，现有的诈骗防御技术仍大多局限于离线下的单模态检测，难以应对当下多模态诈骗的实时性、强交互性挑战。据 GASA 统计，仅 2024 年，我国因诈骗而造成的经济损失超 8971 亿元。

基于日益严峻的诈骗形式背景，本项目以“**跨模态风险穿透分析**”为核心创意，聚焦通话过程中可能存在的诈骗痕迹，突破传统单一维度的防御逻辑，首次将**通话载体、通话行为、通话内容**三者融合大模型技术进行分析。系统重点研究了引导式诈骗检测技术和基于深度学习的伪造检测技术，针对性地设计了**跨模态时序不一致性检测技术、交叉模态情绪一致性检测技术、语义诱导性特征捕捉技术**，在实时通话中同步捕捉伪造痕迹与心理破绽。此外，本作品具备**高可解释性报告的大模型智能助手**。在诈骗识别后，系统将生成详细且易于理解的报告，分析其风险模式与伪造手段，通过图表与可视化工具，使得用户能够轻松掌握防诈信息并进行应对。**风险形象大数据智能推送技术**根据用户的行为特征、历史通话等数据，实现精准化、个性化的反诈教育。**防诈社区**汇集用户反馈与交流，增强集体防骗意识。

本作品易维护、易部署、易拓展，面向多层次用户群体：对普通大众（尤其老年群体），可以提供“无感嵌入”的实时防护，兼容低性能设备；对有关部门，提供有力的辅助审讯功能、舆情谣言监控功能、证据核查功能；对金融机构与电商企业，支持高风险交易前的身份核验与证据链固化；对社交平台，可作为审核系统，拦截伪造广告和风险内容。作品的核心价值在于“技术防御 + 认知赋能”双重突破：既通过多模态分析将诈骗识别准确率大幅度提升，又以可视化风险报告帮助用户理解通信过程中的诈骗逻辑。

当前，全球线上通信、社交平台市场规模超万亿美元，但尚缺一个智能一体化、实用性强的防诈和信息核查系统。而随着《生成式人工智能服务管理暂行办法》、《中华人民共和国反电信网络诈骗法》等政策落地，本作品在线上通信、社交平台、直播电商、司法审讯等场景具备规模化推广潜力，并能以软件平台、系统插件、API 接口等形式快速适配各种客户端。未来，团队将聚焦诈骗，进一步联合公安反诈中心开展场景化迭代，致力于将“一盾当关”打造为数字社会的新型安全基础设施。

---

## 第2章 问题分析

### 2.1 问题来源

#### 2.1.1 新型引导式诈骗领域现状

随着通信技术和数字技术的快速发展，诈骗、谣言等风险内容层出不穷、日新月异，给国家安全、经济发展和社会稳定带来了巨大的隐患。中国互联网协会发布的《2024年中国互联网发展报告》[1]显示，2024年全国共破获电信网络诈骗案件达到50万起，直接经济损失超过近千亿元人民币，全球每年因网络诈骗造成的经济损失以万亿美元计。报告指出，我国诈骗案件的数量和涉案金额逐年上升。尤其是在互联网和移动通信普及的背景下，社交媒体和通信媒体等APP的使用愈加频繁，使得现代诈骗手段变得更加多样和隐蔽。

由于社会发展，现代诈骗相较传统诈骗活动有了许多新特点。互联网的发展，使得诈骗不同于以往依赖于单一模态（如电话/短信）进行传播，而是可以借助音视频通话、线上会议、网络直播、社交平台等多传播平台和音频、视频、文字、图像等多传播媒介进行伪造事实、引导情绪等多模态化诈骗。并且由于生成式人工智能技术的进步，诸如DeepFake等伪造技术为诈骗活动提供加持，诈骗分子可以利用其生成高质量的伪造视频和语音等进行诈骗。例如，通过实时深度伪造等技术，新型诈骗更加**隐蔽化**，诈骗分子可以在与受害者的交互中实时生成伪造视频或音频，而受害者往往缺少这一方面的防备心。同时，大数据的兴起，也为诈骗手段提供了支持，利用大数据和社会工程学技术，诈骗分子可以快速获取受害人画像，通过对受害人兴趣爱好家庭年龄等一系列个人特征的分析，定制出针对某一类人群的个性化诈骗方案，较过去传统的钓鱼短信等，更易成功。也就是说，新型诈骗还拥有**强诱导和精细化**的重要特征。

从诈骗技术手段上，新型引导式诈骗可以大致划分为基于内容生成的**伪造式诈骗**和基于事实哄骗的**诱导式诈骗**。

基于内容生成的伪造式诈骗是指利用深度伪造技术合成高质量的伪造音频和视频内容，进行有针对性和预谋的攻击。这种诈骗行为更加隐蔽，通过冒充身份等手段，极易获取受害者的信任。根据Ponemon研究所2023年的一项调查[2]，在接受调查的553名IT专业人员中，有42%的受访者报告称，公司高管或其家人曾遭到网络犯罪分子的音视频伪造诈骗。这些诈骗导致了知识产权的丧失（占被攻击者的78%）、客户和业务伙伴的流失（66%）以及客户或员工数据的丢失（27%）。此外，世界经济论坛2022年的报告[3]指出，有66%的网络安全从业者在其组织内经历过伪造内容诈骗。

---

基于事实哄骗的诱导式诈骗，通过大数据分析、社会工程学等技术，利用从各种渠道收集到的受害者个人信息来对其定制个性化诈骗话术，使得诈骗过程更具诱导性，诈骗分子更容易与受害者建立信任桥梁。根据360公司发布的《2024年中国手机安全状况报告》，2024年仅是因为诱导式诈骗话术而受骗的人均损失高达36869元，其中90后年轻群体是不法分子从事网络诈骗的主要受众人群，占38.2%。

在现实诈骗场景中，新型诈骗能完美融合伪造式诈骗和诱导式诈骗这两种诈骗技术手段，营造出一场高质量骗局。通过社会学工程为受害者定制点对点的诈骗策略，再通过伪造技术来生成预期的诈骗内容，受害者面对诈骗分子这量身定做、几乎天衣无缝的骗局，仅凭自身往往是无法及时察觉。因此，诈骗分子将能够在短时间内获取受害者的信任并完成诈骗目的。近些年来，缅北诈骗分子便是基于上述所说的新型诈骗，广泛地实施诈骗，除了诈骗钱财，还诱骗中国人前往缅北，涉嫌故意杀人、故意伤害等罪名，对中国造成了严重的经济损失和恶劣的社会影响，成为了一个长期困扰中国人的问题。根据公安部数据，2024年，缅北诈骗案件导致我国公民的直接经济损失估计达到了800亿元人民币以上。此外，据不完全统计，2024年整个缅北诈骗网络可能导致的总损失超3000亿元人民币。

但在实际应用中，诈骗手段也随检测技术而日新月异。其中，伪造技术的不断先进化、实时化，考验着检测技术的及时响应能力。此外，大量变种诈骗话术出现的同时还存在根据人的特征定制个性化诈骗话术。这给诈骗检测防御带来极大的挑战。

### 2.1.2 诈骗的技术方法

我们团队对现代诈骗进行了深入调研，发现绝大多数**诈骗场景**可以归纳为以下两种：**诱导式诈骗**和**强交互式诈骗**。

诱导式诈骗，即通过话术和心理操纵等手段引导受害人上当受骗。诈骗分子可以通过电话或社交媒体联系受害者，假装成受害者的亲友或权威机构的代表，利用受害者的信任和恐惧心理，诱导他们提供个人信息或转账资金。如，犯罪分子可能会假装成银行工作人员，声称受害者的账户存在异常活动，要求受害者提供账户信息以“保护”他们的资金。

另一个典型的诈骗场景是基于内容生成的强交互式诈骗，即利用科技手段生成虚假视频、音频等信息，蒙骗受害人进行诈骗。犯罪分子往往利用深度伪造技术生成高度逼真的伪造视频或音频，冒充受害者的亲友或知名人士，实施诈骗。如，犯罪分子可能会伪造受害者亲友的声音或面部，声称自己遇到了紧急情况，需要立即转账资金。这种诈骗手段由于其高度的逼真性，使受害者难以察觉，从而更容易上当受骗。

---

基于此，我们团队从这两个场景出发进行调研，旨在剖析在这两种场景下，诈骗分子所惯用的欺诈手法，并全面审视当前市场上已有的应对策略与解决方案，以期达到更深刻的理解与有效的防范。

### **(1) 诱导式诈骗**

诱导式诈骗作为一种传统的诈骗模式，其核心在于，**根据部分事实虚构或歪曲事件来欺骗受害者**，使受害者基于错误的信息做出决定或行动。这种诈骗手段存在已久，但其话术和方式却随着时代的发展而不断演变。

上世纪 90 年代，随着互联网的兴起，诱导式诈骗方式迅速更新。钓鱼邮件等诈骗技术被提出，使得诈骗成本大幅降低而效率急剧增高。如，诈骗分子可能通过发送钓鱼邮件，伪装成银行、政府机构或知名企业的官方通知，利用受害者的信任，要求受害者点击链接或提供个人信息以验证账户安全，实则是为了窃取受害者的敏感信息或资金。

到了 21 世纪，随着社交媒体和移动手机的普及，诱导式诈骗的手段更加多样化和精准化。诈骗分子利用社交平台收集受害者的个人信息，如职业、兴趣、社交关系等，然后制定个性化的诈骗方案，通过实时互动和社会心理学技巧，营造紧迫氛围，让受害者来不及仔细思考就做出决定。这种诈骗手段作为传统诈骗的一个里程碑，也是目前最为流行的诈骗方式之一，给诈骗防御手段带来了巨大的挑战。

### **(2) 基于内容生成的强交互式诈骗技术**

基于内容生成的强交互式诈骗核心在于，**诈骗分子利用先进的深度学习与人工智能技术，对目标内容进行高度逼真的重新生成**，这种技术不仅能伪造出以假乱真的视频影像，还能模拟出十分细腻的语音音频，实现了前所未有的欺骗效果。例如，诈骗分子可以通过伪造熟人的人脸和声音，利用受害者对熟人的信任进行诈骗。根据强交互式诈骗的不同依据，我们将其分为基于视频的强交互诈骗和基于音频的强交互诈骗。

#### **(a) 基于视频的强交互诈骗**

基于视频的强交互诈骗中，诈骗分子希望通过技术手段与目标人互换人脸，比如受害者的亲属，以骗取受害者信任。为此，诈骗分子需要使用基于深度学习的人脸伪造技术。

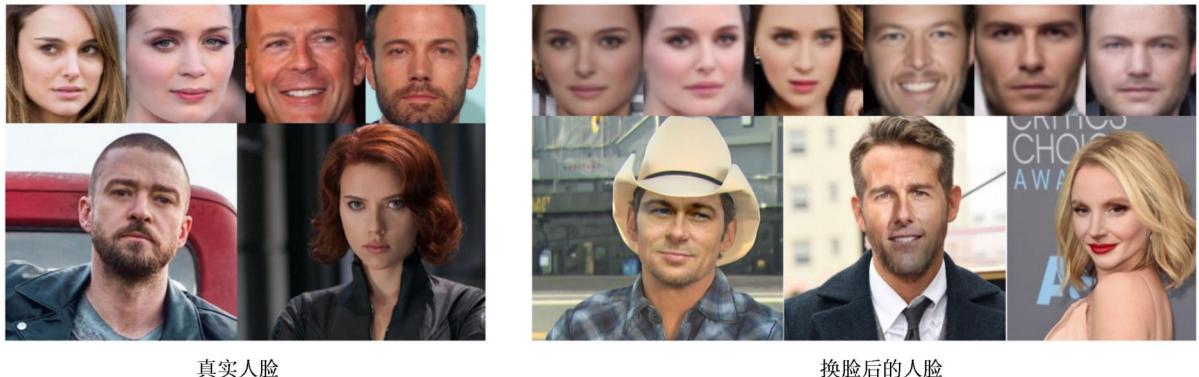


图 2.1 人脸伪造技术使用实例

随着深度学习技术的不断进步，诸如上述的伪造技术层出不穷，显著提升了离线状态下人脸伪造技术的精确度与逼真度，达到了前所未有的高水平。然而，在实时内容生成领域，尽管取得了一定进展，但仍面临技术瓶颈，难以实现高性能且高效能的实时伪造与生成。

2022 年，Yifei Fan 等人提出了 RTDF 框架（Real-Time Deepfake）[4]，该框架旨在帮助用户在直播中进行深度伪造。通过特征提取、热图转换、热图回归和面部融合等步骤，用户可以选择不同的参考面孔创建虚假的非现实面孔。这项技术不仅成功地实现了实时性面部伪造直播，为诈骗分子提供了即时生成并展示高度逼真伪造面孔的能力，还极大地增强了诈骗的迷惑性和可信度，是人脸伪造技术在实时伪造领域的一次重大飞跃，为诈骗活动提供了强有力的技术支持。

与此同时，MyHeritage 推出的 Deep Nostalgia[5] 工具，以及随后问世的 IPLAP[6]、StyleGAN3[7]、StyleSync[8] 等先进技术，不仅极大地增强了伪造视频中人脸的真实感与自然度，使得伪造的视频几乎难以与真实视频区分，为诈骗分子提供了几乎无法被察觉的伪造素材，而且极大地拓宽了视频伪造内容的多样性和表现力，使得诈骗手段更加丰富多变，难以防范。

此外，实时性技术领域的不断突破也为诈骗活动提供了新的助力。首先，硬件上的快速迭代更新，特别是英伟达（NVIDIA）和三星等厂商生成的高性能 GPU，为实时技术提供了强大的运算能力支撑，使得诈骗分子能够更快速地生成和处理伪造内容。其次，实时渲染技术 [9] 和实时处理算法 [10]（如流媒体传输协议 WebRTC[11] 和高性能视频编码和解码技术 [12]）的不断发展，进一步满足了实时无延迟的需求，为诈骗分子提供了更加流畅、自然的伪造视频展示效果，从而更容易骗取受害者的信任。

### (b) 基于音频的强交互诈骗

基于音频的强交互诈骗技术是诈骗分子通过伪造熟人声音，与受害人实时互动，从

---

而获取受害人信任进行诈骗。为提高隐蔽性和实时性，诈骗分子需要实时语音伪造技术。

最早的语音伪造技术是依赖基于规则的语音合成方法，但是这种方法合成的语音具有极强的机械感和不连贯性。

而 2016 年, WaveNet[13] 技术的诞生，更是通过深度神经网络的精妙运用，首次实现了高质量自然语音的生成，如，基于 WaveNet 模型提出了 WaveNet Vocoder[13]，这是一项可以为开发者和研究人员生成逼真音色和自然人声的工具。该项目通过直接预测音频信号的样点，能够捕捉到更细微的声音特征，从而生成更为真实、流畅的语音。这一技术的进步使得诈骗分子能够生成更为逼真的伪造语音，从而进行更具隐蔽性的诈骗。

尽管之前的语音伪造技术以已经取得了一系列辉煌成就，值得注意的是，它们大多仍局限于离线环境，难以满足实时应用的需求。然而，这一局面在 2017 年 2 月迎来了重大转折，百度公司提出了 Deep Voice[14] 这一革命性的文本转语音系统。Deep Voice 不仅保持了高质量的声音输出，而且完全基于深度神经网络构建，更为重要的是，它首次实现了真正的端到端神经语音合成，并成功地将语音伪造技术推向了实时合成的崭新阶段。这一技术的进步使得诈骗分子能够在实时环境中生成高质量的伪造语音，从而进行实时诈骗。

在 Deep Voice 引领实时语音合成新风尚之后，语音合成技术在质量上更是迎来了前所未有的突破。Transfer Learning for TTS[15] 技术的引入，巧妙地将迁移学习的概念融入其中，不仅显著提升了语音合成的效率，更在质量上实现了飞跃。紧接着，2021 年 11 月，YourTTS[16] 横空出世，作为首个集多语言、多说话人及语音克隆功能于一体的语音合成模型，它能够在多样化的语言和声音风格下，轻松生成高质量的语音输出，这也使得诈骗分子能够更轻松地生成多样化的伪造语音，从而进行更具欺骗性的诈骗。

进入 2023 年，语音伪造技术迎来了又一波创新浪潮。HiFi-GAN[17]、NaturalSpeech2[18]、Zero-Shot Multi-Speaker Text-to-Speech with Hidden Unit BERT[19]、Style-TTS2[20] 等技术的相继问世，不仅将合成语音的质量推向了新的高度，还赋予了该技术灵活调整说话人风格、情感乃至无样本生成高质量语音的能力，极大地拓宽了语音伪造技术的应用场景和边界。这些技术的进步使得诈骗分子能够生成更为复杂和多样化的伪造语音，从而进行更具迷惑性的诈骗。

### 2.1.3 诈骗防御挑战

上述伪造技术的发展使得诈骗有了更多更隐蔽的手段，并且在如今的大数据环境下，数据的易获取性也使得诈骗手段个性化、定制化。而各种通讯软件的应用，使得诈

---

骗行为有了即时发生、高度互动的特征。

### **(1) 诈骗方法的多样化、隐蔽化**

诈骗方法多样化指存在多种手段的诈骗。不同于以往只能通过电话、邮件单一地进行逐步诱导的引导式诈骗，现在的诈骗通过伪造生成的高逼真音视频内容和大数据分析得到的人物画像可以设计多种诈骗方法。

诈骗方法隐蔽化指诈骗痕迹难以被觉察。通过深度伪造技术和对话术的精细化处理，诈骗分子可以在强交互式通话中隐藏其真实身份和目的，并使对方无法发现。

### **(2) 诈骗手段的个性化、定制化**

诈骗手段个性化指利用受害者的个人性格、习惯特征，通过社会工程学设计出专门针对受害者个人的骗局。通过社交软件等多种渠道，诈骗分子可以得到受害者的大量个人信息，并利用大数据分析得到其特征画像，以此设法得到受害者的信任。

诈骗手段定制化指根据不同的目标群体或具体语境，设计适用于当前的骗局。通过目标群体的特定需求和具体的情境特点，诈骗分子可以设计出合理的需求导向和情境骗局，精确攻击目标的同时，提高诈骗成功率。

### **(3) 诈骗行为的即时化、互动化**

诈骗行为的即时化指诈骗发生的随机性和无障碍性。通过各种的社交通讯软件，诈骗分子可以在任何时间任何地点，在受害者毫无防备的瞬间，不受阻碍地迅速发起诈骗行为。

诈骗行为的互动化指诈骗行为双方的持续沟通交流。通过各种通讯软件，诈骗分子可以保持与受害者的实时对话，渐渐建立信任纽带，最终以达到诈骗目的。

## **2.2 现有解决方案**

基于前文罗列的各种诈骗技术，为了更好地理解我们项目的针对性，我们调研了应对诈骗的相关现有防御技术。

### **2.2.1 针对诱导式诈骗的识别技术**

诱导式诈骗作为一种传统的诈骗模式，其核心在于通过虚构或歪曲事实来欺骗受害者，使其基于错误的信息做出决定或行动。其诈骗话术和方式一直在随着时代的发展而不断演变，但其诈骗文本的高诱导性的本质并没有改变，也就意味着其往往伴随着与诈骗相关的异常情绪。

#### **(1) 文本语义识别技术**

诈骗分子通常使用固定的套路和话术来诱导受害者上当受骗，其对话中通常包含许

---

多诱导性特征，现在的诈骗文本检测技术主要依赖于简单的关键词匹配方法，这种方法虽然能够识别一些明显的诈骗特征，但面对复杂且多变的诈骗话术时显得力不从心，难以应对各种隐蔽的诈骗手法。生成式大语言模型如 ChatGPT[21]、Llama[22] 等，对文本有较强的理解能力，并可以根据对本文的理解提供合理的解释。据于此，诈骗话术的文本检测有了一定的技术支撑。

2013 年，基于深度学习的识别技术开始出现。这些技术通过学习复杂的文本特征，避免了繁琐的手工特征工程，并能够处理更加隐蔽的短文本诈骗。然而，尽管深度学习显著提升了短文本诈骗检测的能力，但在长文本的处理上仍面临挑战，因为长文本涉及更多的上下文信息和细节，要求更强的分析能力。

2017-2018 年间，Transformer[23] 和 BERT[24] 技术的提出标志着诈骗文本检测进入了一个新的阶段。这些技术引入了注意力机制，使得模型在处理长文本时具有更强的能力，能够更好地捕捉文本中的重要信息。Transformer 的自注意力机制使得模型能够关注到文本中的重要部分，并且更好地理解上下文信息，从而大大提高了对复杂诈骗话术的识别能力。BERT 则通过双向上下文建模，进一步提升了对文本细节的捕捉能力，显著改善了长文本的处理效果。

同年，GPT(Generative Pre-trained Transformer)[25] 推出，2019 年，RoBERTa(Robustly optimized BERT approach) [26] 推出，为诈骗文本检测带来了新的突破。这些预训练语言模型不仅具备了强大的语言理解能力，还通过大规模数据的训练，能够生成更为准确和流畅的文本表示。GPT 的生成能力以及 RoBERTa 的优化技术，为识别更加复杂和隐蔽的诈骗话术提供了强有力的工具。

2020 年 5 月，MMFA (Multi-Modal and Multi-Task Learning for Fraud Detection) [27] 研究推动了多模态学习机制在诈骗话术识别中的应用。这一研究不仅提升了模型在处理复杂诈骗文本中的准确性和鲁棒性，还通过结合多种模态信息，增强了对诈骗行为的综合识别能力。

现代技术在诈骗话术识别上虽取得显著进展，但面对新型诈骗仍显不足。新型诈骗灵活多变，规避现有检测，超出模型训练范围，挑战识别能力。现代技术难以辨别高度个性化诈骗话术真伪，且数据更新滞后于诈骗手段发展。为应对此问题，需要探索新技术，如多模态、强化学习，加强跨领域合作，构建全面识别体系。

我们聚焦于诈骗文本中欺诈和获利的本质特征，不再仅依赖传统的“关键词提取”，而是更加关注“对话”本身的语义特征，并融合大语言模型技术。为了为用户提供更具解释性的证据，我们将文本检测过程分为“语义分类”和“语义解析”两个阶段。在“分类”阶段确定文本类型后，“解析”阶段将针对相关内容进行本质信息的提取，从而设

---

计出更为精准的提示词。通过大语言模型基于这些优化的提示词进行语义解析，能够显著提升输出结果的可解释性，使用户更易于理解和信服。

## (2) 情绪动作捕捉技术

诈骗分子在实施引导式诈骗的时候，由于诈骗本身的“虚假性”和“违法性”，不可避免地会出现情绪的异常波动，比如交流过程中突然心虚或者害怕，而这些都会表现在肢体动作、面部表情或语音情绪上。而利用情绪识别技术能够分析诈骗分子在与受害者交流过程中的语音、面部表情或文本信息中的情绪变化，揭示出潜在的诱导与欺骗特征。当系统检测到异常的情绪波动或模式时，便能及时发出警告，帮助受害者识破骗局，从而有效抵御引导式诈骗的侵害。在这种情况下，情绪识别技术的发展为检测引导式诈骗提供了新的可能性。

目前的情绪识别技术仅限于单模态上，集中在基础语音参数的提取和情绪分类。2000年2月，多模态情绪识别的概念开始出现，Maja Pantic 和 Lijuan Chen 等人开始结合语音和视频数据进行情绪分析 [28]，在情绪识别上开辟新的方向。2010年，深度学习兴起，并进入情绪识别领域。2013年9月，Alex Graves 和 Geoffrey Hinton 等人在深度学习和情绪识别的结合方面进行了开创性的工作 [29]，发表了使用长短期记忆网络（LSTM[30]）进行语音情绪识别的研究，在处理语音信号中的时序信息方面做出重大贡献。同时，在面部情绪识别方面，基于深度卷积神经网络（DCNN[31]）的面部表情识别技术取得了突破，能够从视频数据中提取更准确的情绪特征。此外，Yann LeCun 和 Fei-Fei Li 等人在 DCNN[32] 和计算机视觉领域的技术研究对面部情绪识别起到了重要作用，并进一步推动了多模态融合的情绪识别技术的发展。

在动作捕捉技术上，各种技术层出不穷，其中 yolo 以其轻量化、易部署、精度高而被广泛使用，其可以高效地识别物体，捕捉人体动作，绘制人体关节点图等。且其提供便捷的微调训练策略，易于使用开源数据集或自定义数据集来进行训练。在小规模数据集下，就能达到很高的识别进度。这给高效集成异常动作捕捉技术带来了可能。

而多模态情绪动作识别技术利用多模态数据（如语音、视频、文本等）来提高情绪识别的准确性和鲁棒性。在前期的技术积累下，2020年5月，Jie Yang 和 Chao Zhang 等人在多模态情绪识别方面发表重要成果 [27]，探索基于多模态数据的深度学习模型，这对通过不同模态的信息来提高情绪识别的准确性具有重大意义。2023年，实时的多模态内容情绪识别技术正式大规模地进入应用市场。基于 Ashish Vaswani 等人在 Transformer 模型中的研究成果 [23]，融合了 Transformer 架构的多模态情绪识别模型在处理复杂的情绪表达任务上展现出了卓越的性能，且能够实时应用于多样化的现实场景中，为实时情绪检测提供了强有力的技术支撑。

---

尽管情绪和动作检测技术在学术界取得了显著进展，并已在工业领域得到广泛应用，但在诈骗领域仍存在技术空白，缺乏能够精准、高效识别诈骗行为的检测方法。此外，传统的动作或情绪识别技术难以直接应用于诈骗心理的检测，因为单一的动作或情绪可能具有偶然性或习惯性，无法准确反映真实的诈骗意图。因此，需要设计一种高效的多模态检测系统，从多模态数据维度进行分析，并且在每一个模态上都能精准高性能地进行数据处理，从而更准确、更高效地识别出潜在的诈骗心理。

### 2.2.2 针对伪造内容的诈骗识别技术

自从伪造技术出现后，对应的检测技术便在不断发展。虽然，攻击者可以利用各种先进的伪造技术生成高质量的伪造内容（视频、音频），但是无论如何逼真，其“伪造”的本质并未改变，防御者便可以开发各种检测技术来识别其伪造特征。

#### （1）人脸伪造的检测方法

早在 2016 年，眼睛眨动检测技术问世 [33]，它以其简单高效的特点，有效应对了早期质量较低的伪造内容，但是面对后来棘手的更高级的攻击样本显得无能为力。2018 年 3 月，MesoNet[34] 技术横空出世，能够更深入地识别伪造音频中的微妙痕迹。该技术巧妙融合卷积神经网络（CNN[31]），成功提取伪造特征，并有力证明了 CNN 在 DeepFake 检测领域的巨大潜力。同年 10 月，FaceForensics++[35] 数据集的发布，进一步加强了深度学习在 DeepFake 检测中的应用效果，为后续研究提供了坚实的数据基础。

进入 2019 年，检测技术迈入新的阶段。5 月，“Protecting World Leaders Against Deep Fakes” [36] 研究发布，该研究一改以往只针对视频的单模态检测方式，开创性地提出了多模态伪造检测技术，通过巧妙地将音频与视频数据相融合，对多模态特征进行深入学习和分类，这标志着伪造检测技术的重大突破，推动了多模态数据融合在 DeepFake 检测中的应用。基于这一趋势，多模态 DeepFake 检测技术通过整合音频、视频等信息源，显著提升了检测的准确性和鲁棒性。近年来，随着深度学习与计算机视觉技术的不断突破，多模态检测技术取得了长足发展。2020 年，音视频不一致性检测 [37] 及 LipForensics[38] 等技术的兴起，凸显了面部特征与音频信号在多模态检测中的关键作用。这些技术通过深入分析音频与视频之间的一致性，精准捕捉 DeepFake 视频中的细微差异，极大提升了检测精度。特别是 LipForensics，它利用嘴唇运动与音频信号的同步性，为虚假视频的识别提供了强有力的支持。

近几年来，多模态检测技术的应用流行开来，如 VideoForensics[39]。该技术不仅全面考虑了视频中的上下文信息，还创新性地利用视频的时序特性和上下文进行深度多模态分析。通过细致分析视频帧之间的时序关联，VideoForensics 能够敏锐捕捉视频伪造

---

中的微小异常，从而实现了检测能力与鲁棒性的双重飞跃。这些检测技术上的进步，无疑为我们应对日益复杂的伪造技术提供了强大的工具。

然而，随着相关伪造技术的不断发展，现有的人脸伪造检测技术面对高质量伪造人脸大都力不从心，，面对 lipsync 等多模态领域伪造检测更是无法应对，无法做到高准确率的识别。部分技术虽有高精度的检测技术，但是因为其体量过大、效率低下、针对单一等不足，而无法做到在工业界推广使用。因此，针对目前诈骗领域里多跨模态伪造的特点，我们的技术可以在兼顾单模态的识别技术上，做到对多模态数据的有效、高性能、高精度地检测，易于工业化使用。

## **(2) 音频伪造的检测方法**

诈骗分子可能利用音频伪造技术模拟受害者熟人的声音，因此检测音频伪造是预防诈骗的重要手段。21 世纪初，音频伪造检测主要从主动防御角度进行，如数字水印技术和音频指纹技术等，这些方法虽然可以有效检测音频伪造，但难以实际应用。2010 年后，机器学习技术开始应用于音频检测，从被动防御角度分析语音的生物特征（如语调、节奏、发音习惯等），这是音频检测技术在高效性和实际应用上的重大突破，为后来的研究指明了方向。

随后，GAN 模型、卷积神经网络（CNN）、循环神经网络（RNN[40]）等深度学习技术被引入音频检测中，自动提取音频特征，显著提高了音频伪造检测的精度和效率。通过对抗性训练，这些技术提高了检测算法对伪造音频的识别能力，极大程度上提高了检测的准确率和稳定性，证明了其在音频检测领域的研究价值。

近几年，多模态和实时性等概念在检测领域兴起，音频伪造检测也开始向实时性多模态方向发展。然而，尽管现有的音频伪造检测技术在某些方面取得了显著进展，但它们在应对现代诈骗时仍存在一些不足。首先，传统的检测方法如数字水印和音频指纹技术，虽然能够检测伪造音频，但在实际应用中往往缺乏灵活性和通用性。其次，虽然深度学习技术如 GAN、CNN 和 RNN 在提高检测精度和效率方面表现出色，但它们对伪造音频的识别能力仍然有限，尤其是在面对复杂多变的诈骗手段时。此外，现有技术在实时性和多模态融合方面也存在不足，难以全面应对现代诈骗的多样性和复杂性。因此，现有技术在识别现代诈骗时仍存在较大挑战，需要进一步的研究和改进。针对这一不足，我们需要一个能连续实时性、能处理多模态数据的音频检测数据，以此来应对诈骗领域的伪造音频。

## 2.3 本作品要解决的痛点问题

随着科技的发展，各种诈骗手段和技术日渐先进，诈骗活动也日益猖獗。经过上述调研，本团队发现目前的诈骗防御方式在面对一些先进诈骗手段时表现乏力，同时缺乏统一集成的系统来应对多样化的诈骗行为。其主要存在以下几个痛点问题：

### 2.3.1 防范多样化诈骗手段的性能差

当现有的诈骗检测技术面对不断演变的多样化诈骗手段和不可知的新技术时，往往因为技术落后、更新缓慢或缺乏足够的智能分析能力，以及局限于单一检测模式而显得力不从心，无法及时识别和有效防御新型诈骗行为。因此，在诈骗技术日新月异的背景下，这些技术在实际应用中表现出较低的检测准确率和应对能力，难以满足现代社会对快速、精准识别诈骗行为的需求。

### 2.3.2 抵御实时交互式诈骗的技术少

目前市场上缺乏足够先进的技术来实时监测和防御在交互过程中即时发生的诈骗行为，这主要是由于现有技术无法充分利用上下文信息。实时交互式诈骗往往需要快速反应和高度的实时分析能力，但现有的诈骗检测系统由于算法延迟、数据处理速度不足或缺乏高效的实时监控机制，难以在诈骗行为发生时立即识别并采取措施，从而无法有效保护潜在的受害者免受损失。

### 2.3.3 判断诈骗行为的解释信服度低

在现有的诈骗检测系统中，对于诈骗行为的识别结果往往缺乏足够的解释力和证据支持，导致用户难以完全信任系统的判断。这种不足源于系统在分析诈骗行为时提供的信息不够详尽或逻辑不够严密，使得预警和风险提示的说服力不足，难以令受害者或相关利益方信服并采取相应的防范措施，从而减弱了诈骗防御系统的实际效用和公众对其的信任度。

### 2.3.4 多角度检测系统的可嵌人性弱

面对不断演变的诈骗手段，传统的单一检测技术因角度局限，常忽略关键细节而遗漏诈骗。而多现有的角度诈骗检测系统虽考虑角度全面，但其设计复杂，可嵌人性弱，难以在部署时实现与用户现有系统的良好衔接，且维护成本昂贵，这限制了其广泛应用并影响了诈骗防御措施的效率。

## 2.4 解决的思路

### 2.4.1 作品功能与性能需求

为解决当前诈骗防御存在的痛点缺陷，本团队结合最先进的人工智能技术，设计出一个多模态风险内容检测识别系统——“一盾当关”。该系统以**诱导式诈骗**和**伪造式诈骗**的识别为出发点，能够从音频、视频和图片等多个角度离线或实时地进行诈骗识别，同时，基于《行为心理学》[41]的犯罪行心理学理论，我们在引导式诈骗的识别过程中**创新性地**考虑到了检测对象的情绪角度，旨在结合检测对象的动作、面部表情、语音情感等情绪细节来综合判断检测对象的诈骗嫌疑。最终，系统输出诈骗的识别结果与相应的判断原因。以此来解决当前诈骗防御的问题。

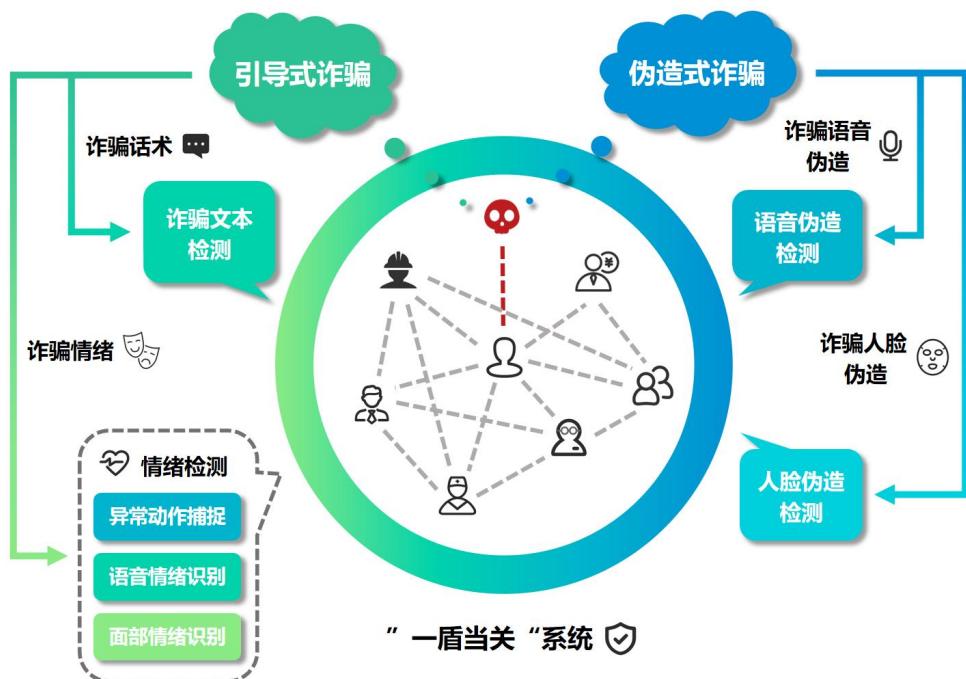


图 2.2 “一盾当关” 系统项目简介

在应对防范多样化诈骗手段、多模态风险内容的问题上，本系统考虑到诱导式诈骗和伪造生成式诈骗等诈骗手段，在图片、音频和视频等多种信息载体上，综合诈骗文本话术、检测对象的情绪、语音或人脸的伪造等多角度进行诈骗识别，多角度多模态地进行诈骗防御，能够在诈骗手段日益多样化的情况下全面地保护用户免受不同形式的诈骗侵害。

在抵御实时交互诈骗方面，系统提供音视频通话过程中的实时诈骗检测。在通话内容上，系统通过判别文本（语音）内容的诱导性事实性特征，进行诈骗特征分类，在大语言模型的强理解能力上根据类别对诈骗话术进行分析，以提供逻辑严密的风险解释；

---

在通话载体上，系统会通过判别载体数据的真伪性直接对通话过程进行诈骗判别；在通话行为上，系统同步地捕捉有诈骗倾向的人物异常动作、面部表情和语音情绪，根据重合的异常情绪时间点定位相应语音文本，通过与文本对比分析，判断检测对象是否存在诈骗心理与诈骗动机。

**在提升判断诈骗行为的解释信服度方面**，系统会结合诈骗检测结果输出识别过程的文字描述，即风险摘要。内容包括伪造、情绪识别、文本诈骗话术的识别，会分别罗列出相应检测目标出现的时间点、检测结果以及系统对检测结果的分析。最终，用户将能清晰直观地了解到系统对通话过程的全方位分析结果，不仅能够察觉到诈骗，还能了解到诈骗分析的原因。如此，不仅能够提升用户对系统诈骗检测的解释信服度，还能够通过诈骗结果分析，让用户了解到诈骗的动机与特征，在下次通话时防患于未然。

**在提升系统可嵌入性、维护性方面**，我们通过采用模块化设计，将系统拆分为多个独立的功能模块，允许用户根据自身需求进行灵活组合或便捷地再开发。同时，系统架构采用了前后端分离技术，便于后期的维护和升级，也提高团队的开发效率。为了优化资源使用，我们设计了一种轻量化架构，使系统能够在低性能设备上也能保持高效运行，这一设计主要考虑到容易被诈骗的群体主要是老年人，而老年人的移动设备一般比较老旧，这确保了即使是在资源受限的环境中，本系统也能稳定工作。

#### 2.4.2 数据集

本团队在开发与测试过程中，使用自研数据集与公开数据集相结合的方式，覆盖诈骗文本、伪造音视频、情绪行为等多模态数据，以全面完善作品功能。通过自研数据针对性增强与公开数据交叉验证，本项目构建了覆盖多模态、多场景的测试体系，确保模型在复杂环境下的鲁棒性，同时严格遵守数据隐私与伦理规范。

##### (1) 诈骗文本数据集

---

表 2.1 诈骗文本数据集说明

项目	内容
核心任务	通话内容：诈骗文本分类与诱导性检测
数据格式	JSON
数据来源	公开网络诈骗案例（如公安通报、社交媒体举报内容），人工模拟生成（基于社会工程学模板设计诱导性话术）
数据获取方式	爬虫抓取公开诈骗案例（过滤隐私信息），团队成员模拟生成高风险对话场景（如冒充客服、虚假中奖）
数据特点	多样性：涵盖 14 类诈骗场景（中奖、虚假招聘、冒充熟人等）；隐蔽性：包含变种话术（如同音字替换、语义模糊化处理）；标注精细：每条文本标注诈骗类型、诱导性强度等级
数据规模	总样本量：8497 条（正常内容 2053 条 + 诈骗内容 6444 条）
评估指标	平均精度 (AP)、均值平均精度 (mAP)

```
01. {
02.   "dialogue_id": "928",
03.   "label": "中奖",
04.   "risk_level": "high",
05.   "text": [
06.     {"role": "诈骗方", "content": "恭喜您！您被选为《幸运星大抽奖》特等奖得主，奖金100万元！请点击链接填写信息领取奖金！"},
07.     {"role": "受害方", "content": "真的吗？需要交手续费吗？"},
08.     {"role": "诈骗方", "content": "只需支付1%公证费即可到账，我们支持微信/支付宝扫码支付，24小时内奖金必到账！"}
09.   ]
10. }
```

图 2.3 诈骗文本数据集样例

## (2) 动态面部表情数据集

### (a) FERV39k 数据集

表 2.2 FERV39k 数据集说明

项目	内容
核心任务	通话行为：动态面部表情识别
数据格式	视频（MP4 格式），每个视频有情感标签
数据来源	公开的视频资料库（社交媒体、影视片段等）
数据获取方式	<a href="https://github.com/wangyanckxx/FERV39k">https://github.com/wangyanckxx/FERV39k</a>
数据特点	多场景：涵盖 4 种主要场景及 22 个子场景；高质量标签： 人工标注保证数据质量；动态视频：适合面部表情识别任务
数据规模	视频片段：38935 个
评估指标	未加权平均召回率（UAR）、加权平均召回率（WAR）



图 2.4 FERV39k 数据集样例

(b) MAFW 数据集

表 2.3 MAFW 数据集说明

项目	内容
核心任务	通话行为：动态面部表情识别
数据格式	视频（MP4 格式），每个视频有情感标签和文本描述
数据来源	电影、电视剧、短视频等来源
数据获取方式	<a href="https://github.com/MAFW-database/MAFW">https://github.com/MAFW-database/MAFW</a>
数据特点	多模态：视频 + 音频 + 文本；多标签：每个视频片段可以有多个情感标签；双语标注：提供英文和中文的情感描述
数据规模	视频片段：10045 个
评估指标	UAR 和 WAR



图 2.5 MAFW 数据集样例

(c) DFEW 数据集

表 2.4 DFEW 数据集说明

项目	内容
核心任务	通话行为：动态面部表情识别
数据格式	视频（MP4 格式），每个视频包含情感标签
数据来源	来自真实世界的动态面部表情视频，收集自电影、电视节目、社交媒体等
数据获取方式	<a href="https://github.com/jiangxingxun/DFEW">https://github.com/jiangxingxun/DFEW</a>
数据特点	动态视频：包含真实的动态表情，适合表情识别任务；多标签：每个视频片段可能有多个情感标签；多场景：视频来源于多种场景（日常生活、影视等）
数据规模	视频片段：11697 个
评估指标	UAR 和 WAR



图 2.6 DFEW 数据集样例

### (3) 静态面部表情数据集

#### (a) RAF-DB 数据集

表 2.5 RAF-DB 数据集说明

项目	内容
核心任务	通话行为：静态面部表情识别
数据格式	图像（JPEG 格式），每个图像有情感标签
数据来源	来自公共面部表情识别数据集、社交媒体和网络收集的面部表情图像
数据获取方式	<a href="http://www.whdeng.cn/RAF/model1.html">http://www.whdeng.cn/RAF/model1.html</a>
数据特点	高质量标签：每个图像由多个标注者进行标注，确保标签的高可靠性；多种表情：涵盖 7 种基本情感（愤怒、快乐、悲伤等）；面部图像：包含高分辨率的面部图像数据
数据规模	图像样本：29672 张
评估指标	AP 和 mAP

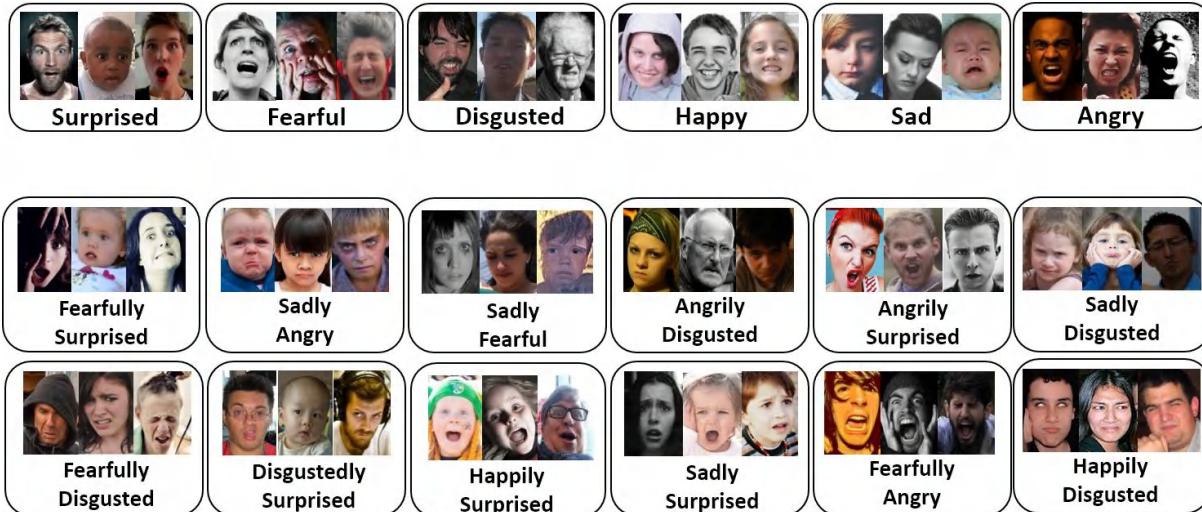


图 2.7 RAF-DB 数据集样例

(b) SFEW 数据集

---

表 2.6 SFEW 数据集说明

项目	内容
核心任务	通话行为：静态面部表情识别
数据格式	图像（JPEG 格式），每个图像有情感标签
数据来源	来自电影中的人物面部表情图像，收集自电影数据库
数据获取方式	<a href="https://cs.anu.edu.au/few/AFEW.html">https://cs.anu.edu.au/few/AFEW.html</a>
数据特点	电影数据：所有数据均来自于电影中的人物面部表情；多样性：包含多种电影类型和场景，表情丰富；高质量标签：每个图像的情感标签由多个标注者确认
数据规模	图像样本：1251 张图像
评估指标	AP 和 mAP



图 2.8 SFEW 数据集样例

#### (4) 动作捕捉数据集

##### (a) Charades 数据集

---

表 2.7 Charades 数据集说明

项目	内容
核心任务	通话行为：异常动作捕捉
数据格式	视频（MP4 格式），每个视频有情感标签和动作标签
数据来源	来自日常生活的家庭视频，包含各种场景和活动
数据获取方式	<a href="http://vuchallenge.org/charades.html">http://vuchallenge.org/charades.html</a>
数据特点	动作识别：不仅标注情感，还包括具体的动作标签；多场景和多活动：视频数据来自家庭环境，场景多样；高质量标签：每个视频片段有精确的情感和动作标注
数据规模	视频样本：9848 个
评估指标	mAP、平均交并比（Mean IoU）

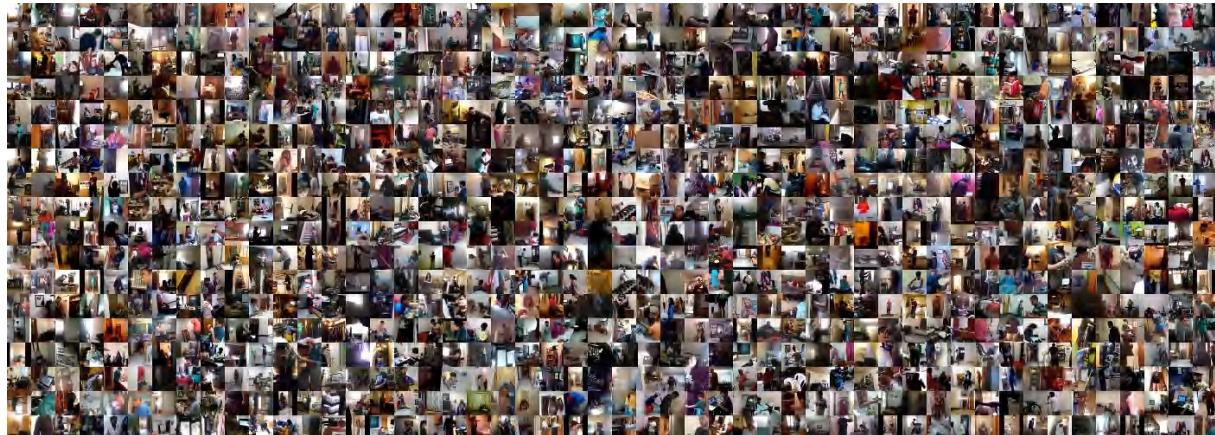


图 2.9 Charades 数据集样例

(b) HVU 数据集

---

表 2.8 HVU 数据集说明

项目	内容
核心任务	通话行为：异常动作捕捉
数据格式	视频（MP4 格式），每个视频有多个标签，包括动作、场景、目标、属性等
数据来源	包含来自多个来源的日常视频，涵盖各种复杂的现实场景，如光线极端、人体遮挡等
数据获取方式	<a href="https://github.com/holistic-video-understanding/HVU-Dataset">https://github.com/holistic-video-understanding/HVU-Dataset</a>
数据特点	多任务标注：涉及多个分类任务，如动作、场景、时间等；复杂场景：包含光线变化、遮挡等挑战；高规模：57 万个视频、900 万个标注，涵盖 3142 个类别
数据规模	实验采用 479,568 个视频，245,868 个标注，739 个动作类别
评估指标	mAP 和 Mean IoU



图 2.10 HVU 数据集样例

## (5) 伪造视频数据集 AVLips

---

表 2.9 伪造视频数据集说明

项目	内容
核心任务	通话载体：视频伪造检测与时序一致性分析
数据格式	视频（MP4 格式）
数据来源	伪造视频由多个主流假脸生成算法（MakeItTalk、DeepFake、Face2Face 等）生成
数据获取方式	团队成员构建
数据特点	经过地标检测、唇部裁剪、多帧检测；部分样本视频实施扰动
数据规模	14500 个伪造视频样本
评估指标	准确率 (ACC)、平均精度 (AP)、假阴率 (FNR)、假阳率 (FPR)



图 2.11 伪造视频数据集样例

## (6) 伪造音频数据集

### (a) LibriSeVoc 数据集

---

表 2.10 LibriSeVoc 数据集说明

数据项	内容
核心任务	通话载体：音频伪造检测
数据格式	音频（WAV 格式），每个音频包含真实和伪造的语音样本
数据来源	从 LibriTTS 语音语料库中提取的真实音频样本，并使用六种神经 vocoder 进行自编码合成生成伪造样本
数据获取方式	<a href="https://github.com/csun22/Synthetic-Voice-Detection-Vocoder-Artifacts">https://github.com/csun22/Synthetic-Voice-Detection-Vocoder-Artifacts</a>
数据特点	六种 vocoder：包含自回归模型、扩散模型和 GAN 模型等三种 vocoder 类型；自编码合成：同一原始样本通过不同 vocoder 重建，生成带有特定伪造特征的音频；多任务学习：用于伪造检测的音频包含 vocoder 识别模块作为附加任务
数据规模	测试使用近 75000 个样本
评估指标	FNR、FPR 和等错误率（EER）

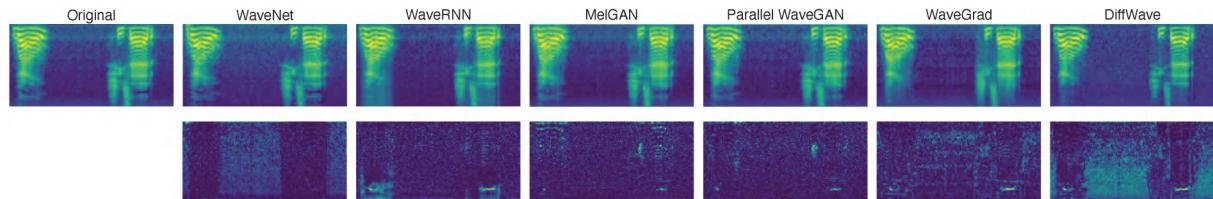


图 2.12 LibriSeVoc 数据集样例

(b) WaveFake 数据集

---

**表 2.11 WaveFake 数据集说明**

数据项	内容
核心任务	通话载体：音频伪造检测
数据格式	音频（WAV 格式），包含生成的伪造语音样本
数据来源	使用多种神经网络模型（如 MelGAN、Parallel WaveGAN、HiFi-GAN 等）基于 LJSpeech 和 JSUT 数据集生成的语音样本
数据获取方式	<a href="https://github.com/RUB-SysSec/WaveFake">https://github.com/RUB-SysSec/WaveFake</a>
数据特点	多种生成模型：包括 MelGAN、Parallel WaveGAN、Multi-Band MelGAN 等；多个语音数据集：LJSpeech、JSUT 和全 TTS 流水线生成的数据；高质量音频：生成的音频具有与真实语音相似的特征，适用于 DeepFake 音频检测研究
数据规模	总共约 175 小时的音频数据，包含 104885 个生成的音频片段
评估指标	FNR、FPR、EER

(c) ASVspoof 2019 数据集

**表 2.12 ASVspoof 2019 数据集说明**

数据项	内容
核心任务	通话载体：音频伪造检测
数据格式	PCM 波形，经过 Flac 压缩，无电话或移动设备编解码器
数据来源	VCTK 基础语料库，包含来自 107 名说话人的语音数据（46 名男性，61 名女性）
数据获取方式	<a href="https://www.asvspoof.org/database">https://www.asvspoof.org/database</a>
数据特点	多种欺骗攻击类型：TTS、VC 和回放攻击；两种评估场景：逻辑访问（LA）和物理访问（PA）；多样的攻击算法：训练和开发集包含已知攻击，评估集包含未知攻击
数据规模	近 25GB 大小
评估指标	FNR、FPR、EER

---

## 第3章 技术方案

随着互联网技术的快速发展和信息传播方式的多样化，诈骗手段日益复杂，传统的防诈骗方法逐渐难以应对新型的引导式诈骗，尤其是那些结合深度伪造技术和大数据分析的诈骗方式。因此，本系统“一盾当关”旨在通过多模态风险识别技术，实时监测并有效识别各种新型引导式诈骗行为，保护用户免受欺诈损失。

### 3.1 诈骗风险内容识别系统设计方案

#### 3.1.1 设计思路

目前，谣言等风险内容不断涌现，诈骗横行且其手段多样化、技术先进化，对社会安全、公民安全有着严重的危害，但现在仍缺少一个可以高效灵活地检测多模态风险内容的大模型系统。

针对这一需求，本团队基于调研诈骗过程中发现的实际问题，研发了一套针对性的防御识别系统“一盾当关”。我们从受害者的角度出发，全面深入分析通信过程，对受害人可能接收到的所有诈骗源进行风险排查，力求全方面地检测诈骗。基于这一思路，系统将从以下三个方面进行诈骗检测。

**通话内容：**包括实时通信的交谈内容和短信等离线文本内容。我们将分析对话中的语言、关键特征词、语境等，以识别潜在的诈骗痕迹。

**通话载体：**包括视频和音频。我们将利用图像和声音的特征，检测是否存在伪造、篡改或虚假的内容。

**通话行为：**包括通话中检测对象的动作、面部表情和语音情绪，同时结合通话内容的语境进行分析判断。例如，某人是否慌张、是否有不寻常的举止，都可能暗示诈骗风险。

围绕“实时高效识别诈骗”的中心目标，该系统从以上三个方面展开算法设计，并结合考虑展示、业务、数据三个层面的进行功能设计。系统能够做到精确判别通话内容的真实性、鉴别通话载体的真伪性、分析通话行为的心理，并进行风险点定位，最终提供风险摘要内容，做到有效规避受骗风险。

#### 3.1.2 技术路线总览

系统结合了视频、音频、文本及行为数据，能够多角度全面地进行诈骗检测。这种多模态的检测手段使得系统能够同时验证通话内容的真实性、通话载体的真伪性以及行为中的心理暗示，大大提升了识别的准确性。尤其在面对复杂的诈骗场景时，系统能够

灵活应对，通过多维度的数据分析，消除了单一模态可能存在的偏差，使得诈骗识别更加精准，误判率大幅度降低。

此外，系统具备强大的鲁棒性和适应能力，能够处理不同类型和复杂度的诈骗手段，无论是面对新兴的诈骗模式，还是复杂的跨模态攻击，系统都能保持较高的识别准确性。系统能够不断学习和优化其防护策略，以应对不断变化的诈骗行为，确保能够有效识别并防御未来可能出现的威胁。此外，系统将基于用户数据来刻画风险形象，结合大数据，针对防御薄弱点推送相应诈骗案例和教育资料，并涵盖用户未接触过的诈骗手法，全方位提升用户防诈骗意识。

本系统的核心算法包括三大算法：通话载体维度的基于跨模态时序不一致性的伪造检测算法、通话行为维度的基于交叉模态情绪一致性的诈骗心理识别算法和通话内容维度的基于文本诱导性特征捕捉算法，其技术总览如下：

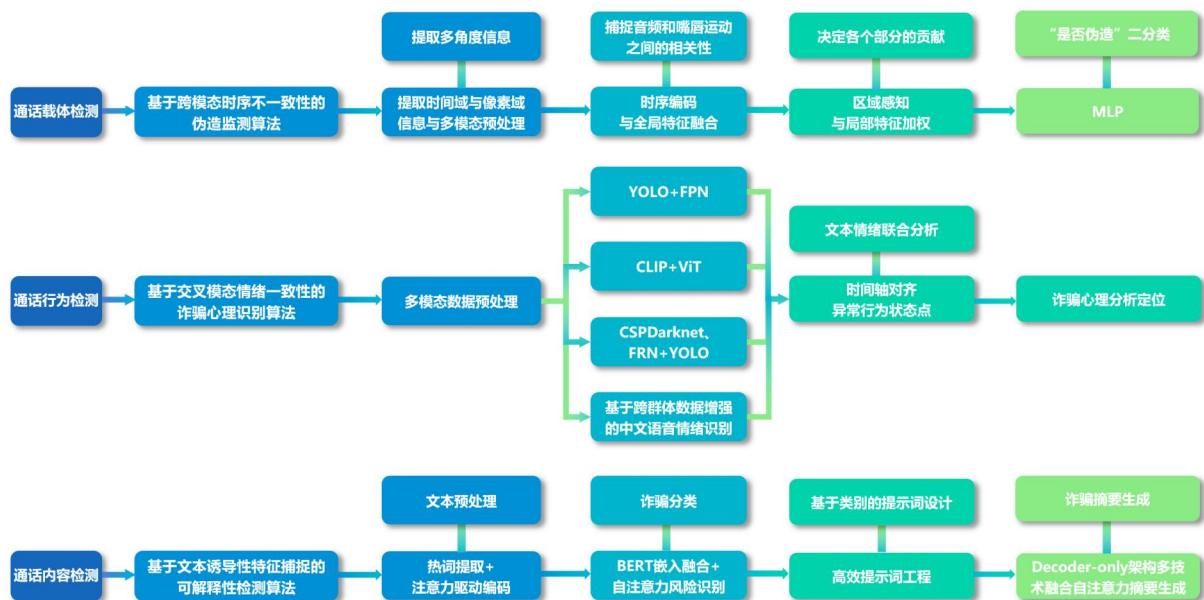


图 3.1 核心技术路线图

- 通话内容事实判别部分主要使用音频转换器与文本检测模块，通过使用语音识别技术，将音频中的语言信息转化为可处理的文本内容。同时，系统基于深度学习模型训练，学习并识别诈骗话术的特征，对输入的文本进行风险评估，识别潜在的诈骗模式和行为。
- 通话载体真伪检测部分使用伪造检测模块，该模块通过时序不一致性分析，识别深度伪造内容，如 Deepfake 和 LipSync 技术伪造的音视频。系统通过分析音视频信号的同步性差异，检测是否存在不自然的修正或伪造痕迹，从而确保通话内容的真实性。
- 通话行为心理测试部分使用语音情绪识别与面部表情识别模块和行为检测模块，语

音情绪识别模块通过情感计算分析语音的语调、语速特征，判断说话人的情绪变化，揭示通话中的心理状态。面部表情识别则通过图像处理技术分析通话者的面部表情，推测其情绪波动，进一步判断是否存在潜在的诈骗行为。行为检测模块不仅仅分析语音和视频，还结合行为学的分析，通过监控动态面部表情和异常动作捕捉，识别诈骗者的异常行为模式。这一模块确保能够识别那些可能通过异常肢体动作或语气变化引发的诈骗风险。

### 3.1.3 系统功能设计架构

项目总体功能设计结构图如图3.2。

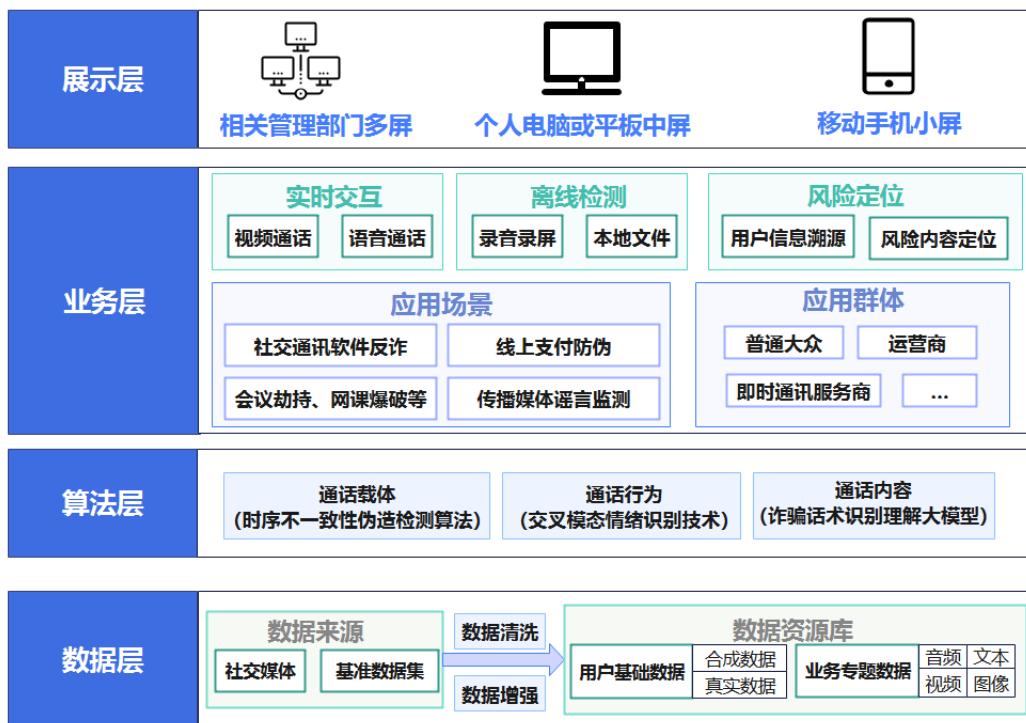


图 3.2 总体功能设计架构图

其中，展示层可以位于多个操作平台上，如管理部的多个屏幕、个人电脑或平板中屏和移动手机小屏。通过易部署性和可嵌入式设计，诈骗防御系统可以保证在满足能移植到多个操作平台的同时，维持其高效的检测性能。

业务层面向实时或离线下的多种任务需求、多种应用场景和多种应用群体。为满足最基本的通话功能，系统开发可实时交互的视频通话和语音通话，允许双人或多人实时通话；为满足录音录屏文件、本地文件的诈骗检测功能，系统在注重实时检测的基础上，依然保留离线检测功能，帮助用户检测文件内容的诈骗痕迹；为满足各种实际的应用场

---

景和应用群体的特定功能，本系统采用模块化设计，保证良好的可拓展性，满足再开发需求。

算法层可以处理通信过程中的多模态数据，从通话载体、通话行为、通话内容三个角度分别设计伪造检测算法、情绪识别技术、文本检测模型（诈骗理解大模型），做到对诈骗痕迹的全方位检测。

数据层可以用来进行系统训练和测试模型。其中的数据来源于各个模型的基准数据集和社交媒体、通讯软件等真实场景，并通过数据清洗和数据增强对数据集进行处理，以提高模型的性能和系统的可靠性。

总之，在系统功能的设计方面，团队致力于为用户提供在实时场景或离线情况下的通话内容、通话载体、通话行为的全方位检测服务。在满足用户通话需求的基础上，能高效地检测多模态内容中的诈骗意图和风险内容。然而，在系统实现过程中，我们主要面临以下挑战：

**①功能方面：应对攻击方式的多样性和不确定性。**结合了深度伪造和大数据分析等技术，诈骗手段可以不断迭代更新而呈现多样化。诈骗分子可以通过单独或融合使用伪造音视频内容、定制诈骗话术等多种诈骗技术来进行诈骗过程的实施，使得诈骗攻击方式多样化和不确定化。这极大程度上考验着系统诈骗检测功能的合理设计和优秀性能

**②系统方面：确保前后端交互方式的高效性和系统的易维护性**系统采用前后端分离架构，包含多个模块组件，涉及到多个模态的数据传输、检测和联合分析，并需要对用户、运营商等多个使用者进行合理安排功能权限，对前端合理交互提出要求。如果交互方式不合理，将导致系统响应延迟，无法及时监控并检测通话内容，进而不能及时向用户发送诈骗警告，严重影响防诈骗效果。

**③工程方面：实现开发技术模型的迁移性和兼容性。**本系统采用了多种检测技术，包含很多依赖库和使用工具，需要协调保持所有检测技术的正常运行。并且当移植到不同的操作系统和软件平台时，因为不同硬件和软件环境具有各自的特性和限制，导致系统不兼容而无法正常运行。若无法保证模型的迁移性和兼容性，将会严重限制系统的应用场景和发展潜力。

为了解决上述三种挑战，团队合理地设计系统架构，以保证各项功能和作用的正常运行。系统架构图如图3.3所示：

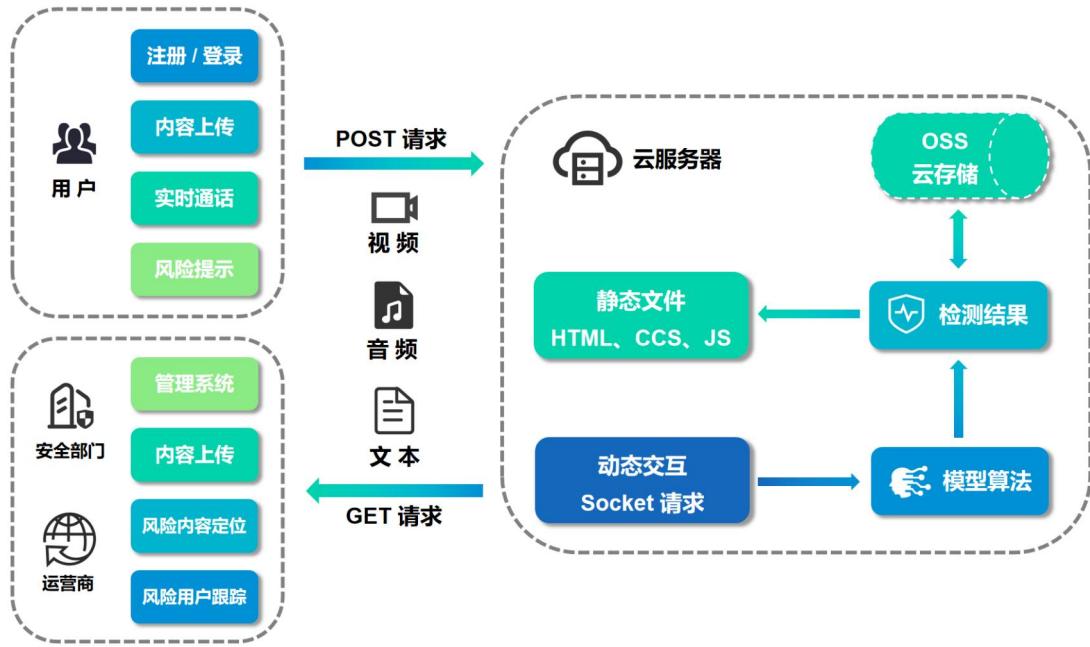


图 3.3 “一盾当关” 系统架构

用户对前端发送 POST 请求，将通话过程的数据发送到服务器，调用后端算法对数据进行处理，并将诈骗检测结果返回到前端页面展示。特别地，系统对不同身份的用户返回不同的检测结果。如运营商、安全部门等可以在得到普通用户的权限功能外，可以额外得到其他权限功能，如风险用户信息、风险用户追踪。

而系统的 Web 服务器采用 NGinx[42] 和 uWSGI[43] 技术，能够实现高效、安全和可扩展的 Web 服务架构。Nginx 负责处理前端请求、静态文件和安全管理，uWSGI 则专注于运行 Python 应用和处理动态内容。

#### 3.1.4 系统设计

本系统从实际的诈骗场景出发，围绕通话载体、通话内容、通话行为三个方面展开系统模块设计，其中包括通话载体的视频伪造检测、音频伪造检测，通话内容部分的音频转换器、诈骗文本检测，通话行为的异常动作捕捉、面部表情识别、语音情绪识别。

##### (1) 面向通话载体的视频真伪鉴别

人脸替换和唇音同步技术是伪造视频中的两种最为主流的方法，而本项目的视频真伪鉴别可以准确地检测上述两种伪造内容。LipFD[44] 检测技术是视频真伪鉴别的核心部分，不同于其他检测技术都仅仅关注单帧与音频信号的对齐，本检测技术聚焦于帧与帧之间的联系，关注连续帧的时序性对齐。通过接收来自前端的视频数据，研究连续帧间的唇部运动和音频信号，分析其时序不一致性而判别视频真伪。最后输出判定结果：真或假，伴随一个置信度。

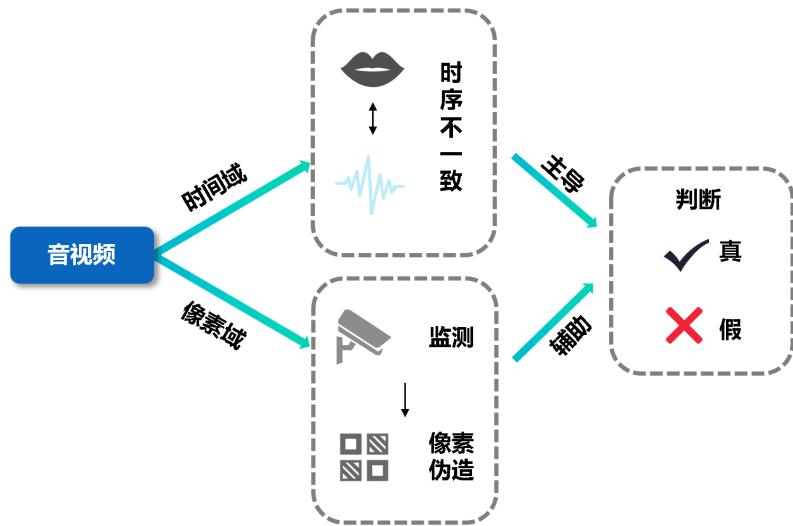


图 3.4 视频真伪鉴别流程图

## (2) 面向通话载体的音频真伪鉴别

诈骗的实际场景中，音频伪造内容泛滥，是伪造式诈骗的基础，而音频真伪鉴别专门来检测伪造语音中的声码器产生的特征部分，即声码器伪影。不同的声码器会产生不同的伪影，我们通过训练让声码器检测器“记住”这些特征，在应用时接收来自前端的音频内容并进行检测，会识别出其中的声码器伪影部分从而判别输入音频为假。工作流程如图3.5：

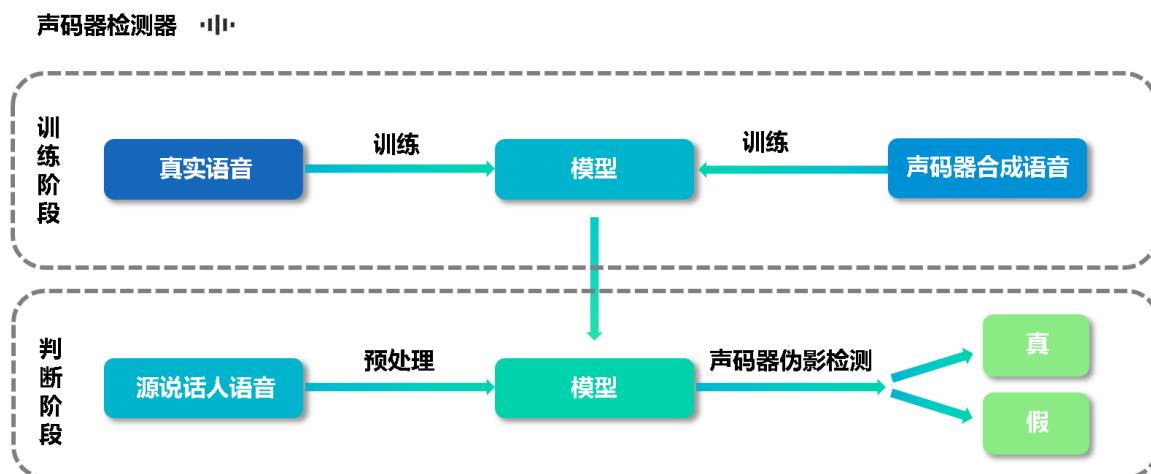


图 3.5 音频真伪鉴别流程图

## (3) 面向通话内容的音频形式转换

诈骗分子一般通过音频形式传递信息。由于大语言模型无法直接处理音频内容，我们需要先将接收到的音频内容转换为文本，后续对文本进行分析处理。首先，对语音内容进行生成和热词处理，提取出关键热词，并对其进行嵌入编码。接着，通过偏置解码器处理这些编码信息，并使用注意力分数过滤机制，确保信息的准确性和相关性。最终，系统输出对应的文本内容，从而实现对音频内容的有效分析和处理。工作流程如图3.6：

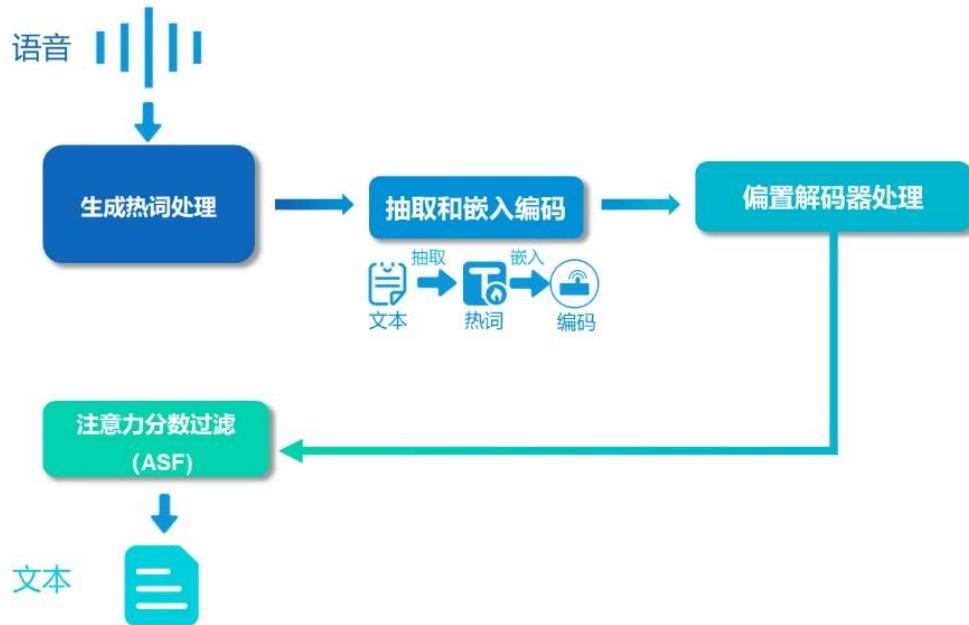


图 3.6 音频形式转换流程图

#### (4) 面向通话内容的诱导文本判别

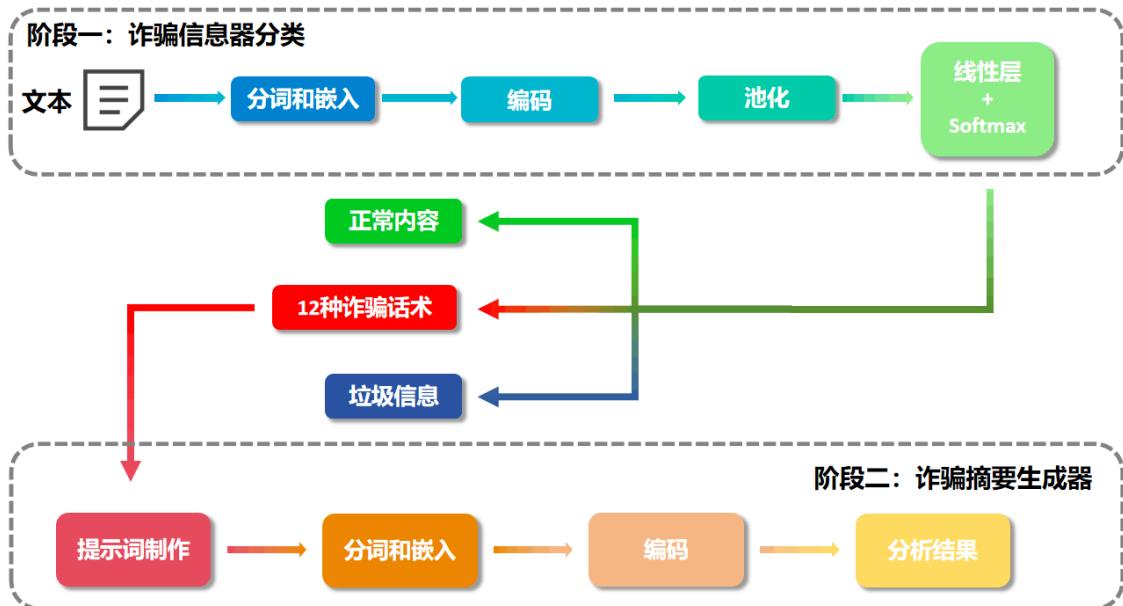
诈骗分子通过诱导性话术进行诈骗的案例广泛存在，已成为主流诈骗手段之一。这些话术通常通过巧妙的语言技巧和心理操控，诱导受害者做出不利于自己的决定。为了有效应对这一问题，我们采用了诱导文本判别技术，能够检测出隐藏在文本内容中的诈骗话术。

在接收到音频转换后的文本内容后，我们设计的诈骗摘要生成器会根据可疑文本的类别提供详细的风险解释。处理问题时，我们采用了两轮预测的方法，以确保检测的准确性和可靠性。

首先，我们对可能的诈骗信息进行了预分类。在第一轮预测中，我们会让文本分类器预测文本对应的类别下标。这一步骤可以明确是否存在诈骗行为以及其所属类别，抓住该诈骗文本的本质类别特征，同时有利于设计提示词，鼓励大模型生成更具解释性和准确性的输出内容。如果检测到存在诈骗行为，文本将进入第二轮预测。

在第二轮预测中，我们会基于第一轮得到的诈骗类别设置提示词，指定让摘要生成器根据诈骗信息所属类别进行详细解释。采取两轮预测的原因在于，这样可以更加精准

地判断文本的类别，并制作出更加高效的提示词，使得大语言模型能够给出更加精确的回答。通过这种方法，我们能够有效地识别和防御诱导性话术诈骗，保护潜在受害者免受损失。工作流程如图3.7：



## (5) 面向通话行为的异常动作捕捉

在诈骗过程中，诈骗分子可能由于“心虚”或“紧张”，会在通话过程中不自觉地表现出一些异常动作，例如眼部斜视、手摸鼻子等。这些异常动作往往是诈骗心理的外在表现，因此，通过捕捉这些异常动作可以有效识别诈骗行为。

当接收到视频后，我们的视频异常动作捕捉系统会对视频的每一帧进行详细的异常动作捕捉。系统会分析每一帧中的动作，预测其类别，并以秒为单位将结果返回给 App 系统。为了防止误判，我们设计了算法来忽略连续时间内的异常动作。例如，如果对方只是长时间将手放在头上或脖子处，这并不属于异常动作。最终，系统会根据每一帧所处的时间（秒）以及其对应的动作，输出一个动作-时间（秒）列表。这种方法不仅能够准确捕捉到诈骗分子的异常动作，还能有效减少误判，提升系统的准确性和可靠性。工作流程如图3.10：

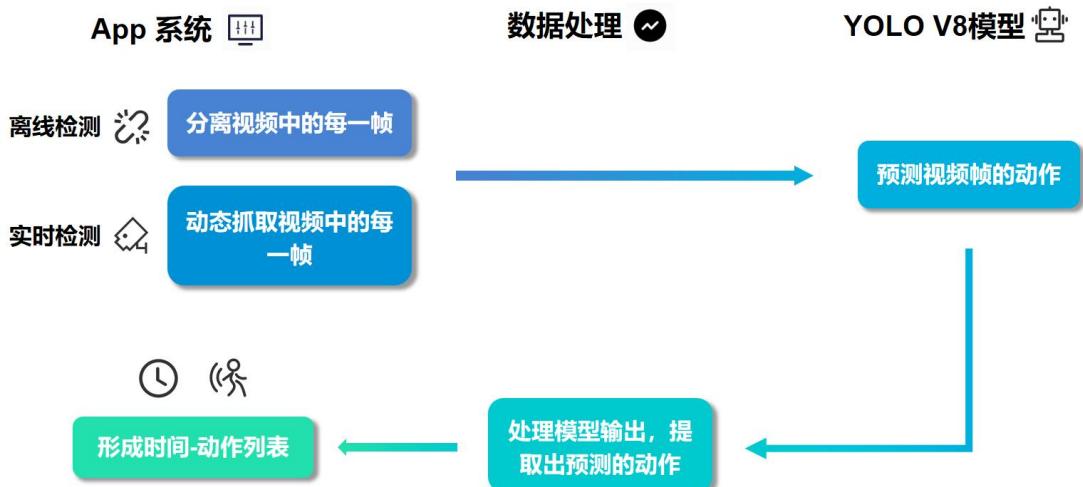


图 3.8 异常动作捕捉流程图

#### (6) 面向通话行为的视音情绪识别

与上文同理，诈骗分子也会表露出一些异常面部表情和异常声音情绪。视频情绪识别会检测带有诈骗心理倾向的异常表情，其技术核心是“动态的情绪识别”，结合某一视频帧的前后帧进行时间维度上的特征分析。接收视频后，将视频裁剪成帧，然后针对帧通过结合其前后帧进行动态情绪预测。而语音情绪识别会检测带有诈骗心理倾向的异常语音。接收音频后，通过分析音频信号的各种特征，如音调、语气、节奏等，识别出音频内容中所表示的情绪状态。两种技术都会对异常情绪进行时间点定位，并输出异常情绪时间点。视频情绪识别工作流程如图3.9，语音情绪识别与其类似。

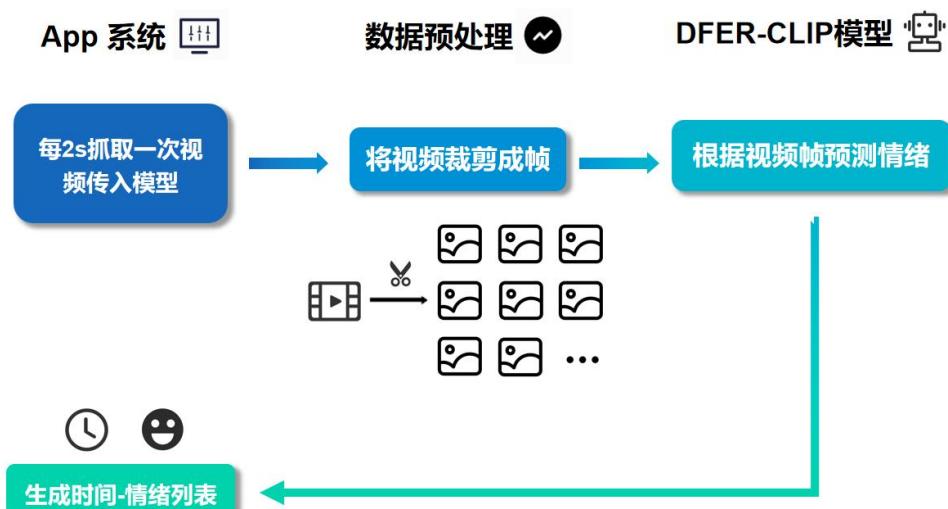


图 3.9 视频情绪识别流程图

#### (7) 数据管理与应用

在多模态诈骗检测系统中，数据库模块通过 Django 框架实现，负责存储和管理诈骗信息、用户上传的音视频文件及处理结果，其中音视频数据存储在外部存储系统（阿里云 OSS），元数据（如文件路径、类型、大小等）存储在 Django 数据库中，使用 ORM 模型关联文件与用户信息，确保数据安全与高效检索。

其次，系统通过 RESTful API 与外部模块（如云大模型）进行交互，处理后的音视频数据结果返回并存储在数据库中，更新文件的处理状态。为了提高响应速度，系统使用 Django RQ 等异步任务队列管理音视频处理，前端无需等待后台任务完成。

值得一提的是系统通过分析用户历史数据，可以识别防诈骗薄弱环节，构建精准的风险画像，并推送相关诈骗案例和教育资料，帮助用户提高防骗意识，逐步强化反诈能力，其运行效果图如下：

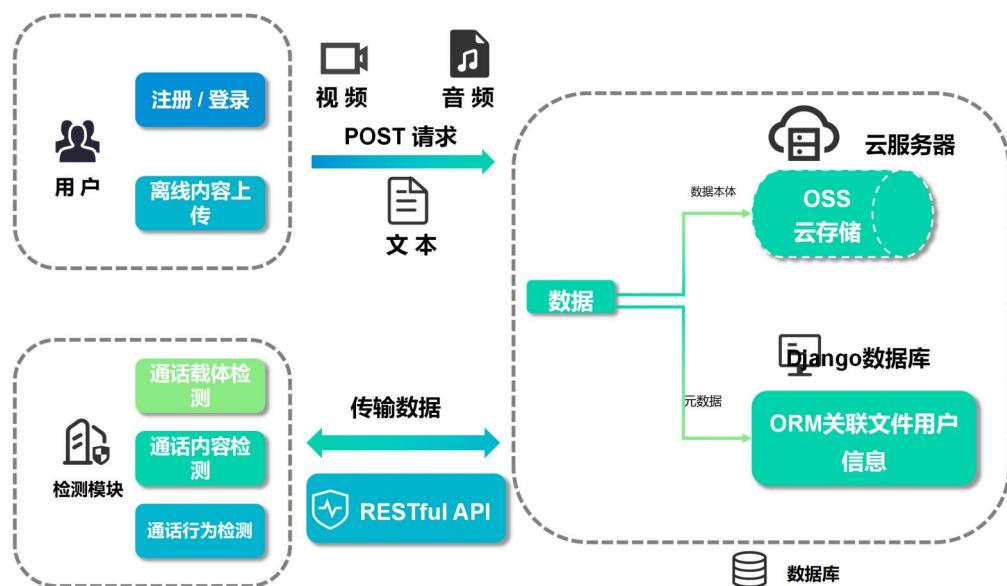


图 3.10 数据库运行效果图

## 3.2 关键技术与原理

本项目围绕通话过程中的载体、内容、行为这三个层面来研发系统的诈骗检测功能，并设计了以下三种关键性技术，以满足我们的功能需求。

### 3.2.1 基于跨模态时序不一致性的伪造检测算法

随着深度学习的发展，人脸替换技术（Faceswap）、唇同步技术（LipSync）也不断发展中更加先进性，并因其可以生成高质量的人脸伪造内容而被引入到诈骗领域中。前

---

者是将诈骗人的面部替换为受害者信任对象的面部实施诈骗，而后者则使用受害者信任对象的真实身份信息，视觉层次上仅仅修改了唇部区域。这类技术的生成伪影难以被人眼辨识，对唇形和口部细节进行高度还原，使得伪造的唇部动作看起来非常自然且与语音完全匹配，这导致现有的检测方法大多对其失效。因此，针对性研发一种可以检测 LipSync 伪造内容并同时兼顾检测一般伪造内容的技术成为了当前研究的一个重要领域。

研究发现，音频信号具有连续性，天然地无法对齐离散的视频帧，这导致 LipSync 伪造技术合成的虚假视频或者图像具有时序上的伪造痕迹。我们团队进一步验证了这种时序上的差异，与导师实验室合作设计了能提取音视觉模态中内在不一致性的双头检测结构，并通过捕捉嘴唇和头部运动之间的关系来模仿人类的自然认知，以提高检测准确性。

此外，为了保证能有效地部署到落地应用，团队对模型进行轻量化处理，从而较少模型的计算开销，使其可以嵌入到大多数系统和平台中。

接下来，我们将从模型设计的两个重要方面：时序不一致性捕捉和轻量化处理，展开详细介绍，凸显模型的技术核心。

### **(1) 基于时序不一致性的差异捕捉**

从时序角度上，我们观察对齐的连续视频帧和音频频谱，如图所示，可以发现音频谱中的能量变化与唇部运动在时序上具有高度相关性。在真实模式下，说话人的唇部运动与其头部姿态和话语内容紧密交织，形成了自然而流畅的统一体，这些微妙的身体语言与说话的时机和语境相辅相成。在伪造模式下，合成的音视频在时间同步上难以隐藏固有的不一致性，其会体现在嘴唇的运动幅度、频谱的明暗变化等，而这些不一致性便凸显了自然嘴唇运动与人工生成嘴唇运动之间的微妙差异，为检测技术的研究提供了宝贵线索。



图 3.11 真假样本在梅尔频谱上的差异

为了让模型能提取上述时序特征，我们需要对所有数据样本进行预处理。首先使用预处理算法构建嘴唇居中且包含完整头部的视频，随后将其与音频频谱进行对齐以获得时序性样本，最后将其裁剪成不同区域来表征不同大小的特征范围。具体算法的伪代码见算法1。

---

#### Algorithm 1 数据预处理

---

```

1: 输入: 一个视频文件  $v$  和一个唇部特征映射  $lips$ 
2:  $landmarks, boundingBox \leftarrow \text{LandmarkDetector}(v)$ 
3:  $videoCenter \leftarrow \text{getLipCenter}(landmarks[lips[0] : lips[-1]])$ 
4:  $v \leftarrow \text{fitCenter}(v, videoCenter)$ 
5:  $v \leftarrow \text{cropVideo}(v, boundingBox, 1.5 \times boundingBox.shape)$ 
6: save  $v$  to dataset
7: set window length  $T$  to 5 and sample extension rate  $r$  to 10
8:  $spec \leftarrow \text{getSpectrum}(v.audio)$ 
9:  $startingPoints \leftarrow \text{random}(0, v.length, r)$ 
10: for  $s$  in  $startingPoints$  do
11:    $raw \leftarrow \text{concat}(spec[s : s + T], v[s : s + T])$ 
12:   add  $raw$  to  $rawSamples$ 
13: end for
14: crop  $rawSamples$  based on head, face, lip to get  $\{c_h, c_f, c_l\}$ 
15: 输出: region samples  $\{c_h, c_f, c_l\}$  and temporal samples  $rawSamples$ 

```

---

## (2) 基于双头检测的技术结构

为了充分提取音视频时间流中的内在不一致性，本团队开发设计了一个基于双头检测的技术结构，从时间域与像素域同时进行伪造检测，如图所示。我们的框架分为四部分，首先是时序编码模块，建模唇部运动和音频关系，然后是全局局部编码器，对视频帧以唇部为中心进行不同范围的裁剪，对三种不同大小的区域与全局特征进行联合编码，然后是局部感知模块，使模型对不同大小面部区域赋予权重，决定每个区域对于最终判别的器贡献，最后判别器使用多层感知机来判别真伪。

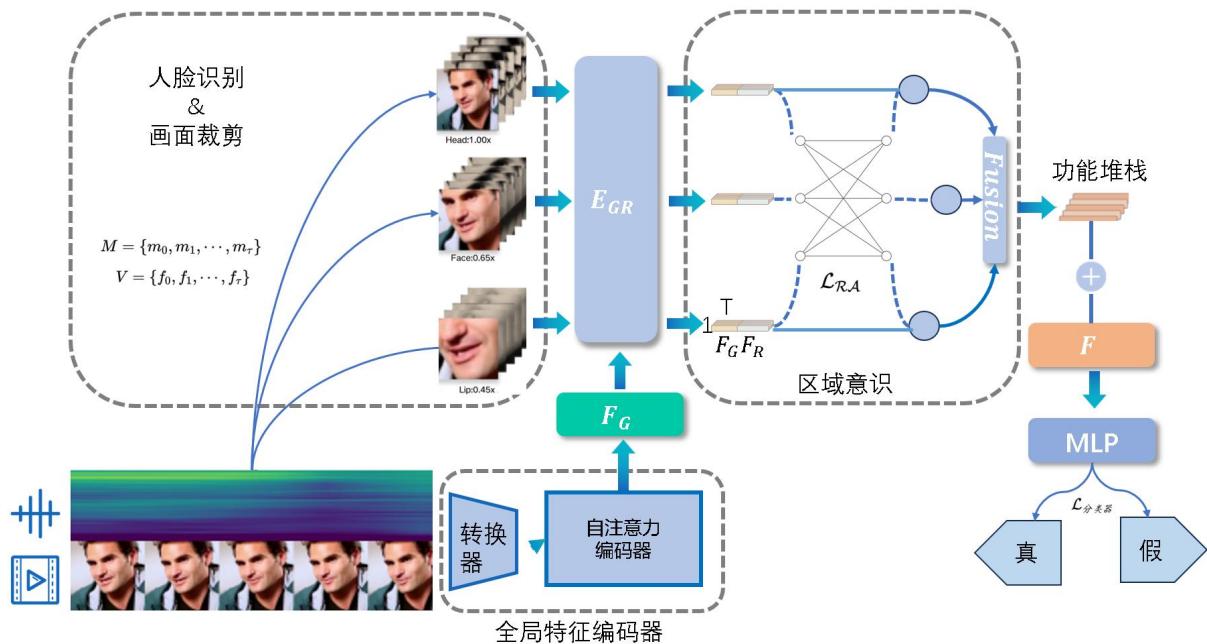


图 3.12 算法整体框架

我们的模型核心部分如下：

### (a) 全局特征编码器 (Global Feature Encoder)

全局编码器将嘴唇运动与频谱间的关联看作 NLP[23] 领域中单词与语句的关系，通过 VisionTransformer[45] 编码时间特征，捕捉音频和嘴唇运动之间的相关性。我们使用预训练的 CLIP 编码器 [46] 的最后一层 ViT-L/14[45] 作为编码器。因为该模型的表征能力强，能够准确地分配注意力到感兴趣的区域。原始图像定义为  $I$ ，我们将卷积层表示为 Conv，它将图像卷积到  $224 \times 224$ ，然后由 ViT 进行嵌入，得到全局特征  $FG$ 。同时将原始的视频帧裁剪为头部、脸部、嘴唇，即  $ch$ ,  $cf$ ,  $cl$  三个序列以表征不同范围的特征。

$$FG = ViT(Conv(I; \theta_{Conv})) \quad (3.1)$$

---


$$\{c_h^N, c_f^N, c_l^N\}_i = Crop(I, \{1.0, 0.65, 0.45\}), \quad i \in \{0, 1, 2\} \quad (3.2)$$

### (b) 局部感知器 (Region Awareness)

在不同的尺度上动态调整模型的注意力，利用了来自不同大小区域的特性，从而使我们的模型能够有效地捕捉 DeepFake 中的显著变化和 LipSync 中的微妙调整。对于每类裁剪  $c \in \{c_h, c_f, c_l\}$ ，区域特征被定义为  $F_R = E_{GR}(c, F_{G, GR})$ ，我们希望这个模块可以关注不同裁剪区域中信息最丰富的部分，其中  $c_l$  对应唇形， $c_h$  和  $c_f$  分别对应头部姿势、面部表情。

由于 LipSync 通常仅在嘴部略有改动，非监督模型可能无法学习到适当的表示。因此，我们进一步引入了一个区域感知模块，它是一个经过修改的全连接层，后面跟一个 sigmoid 函数，考虑到区域切片中的各子区域以及区域特征与其相关全局上下文之间的关联，我们为它们赋予不同的权重，权重表示在最后的判别中所拥有的贡献程度。权重公式如下：

$$\omega_{c_j^i} = RA([F_G | \{F_R\}_j^i], \theta_{RA}), \quad c_j \in \{c_h, c_f, c_l\} \quad (3.3)$$

$\omega_{c_j^i}$  表示  $c_j$  中第  $i$  个特征的权重， $\theta_{RA}$  是局部感知模块的参数。得到权重后，将所有的特征向量按照他们对应的权重进行融合，在最终的融合特征向量中，高权重的特征将会占据更高的比重，进而主导最终的判断。

$$F = \frac{1}{T} \cdot \frac{\sum_{i,j} (\omega_{c_j^i} \cdot [F_G | \{F_R\}_j^i])}{\sum_{i,j} \omega_{c_j^i}} \quad (3.4)$$

无论是 DeepFake 还是 LipSync，其修改区域全都集中在头部的某个区域中。为了让模型能够准确的关注到相关的部分，我们设计了一个 Region Awareness Loss，用来约束模型的注意力。该损失函数通过最大化面部和唇部的权重来人为约束模型，将其关注的重点集中在更高概率被修改的区域上。该损失函数的定义如下：

$$\mathcal{L}_{RA}(\theta_{GR}, \theta_{RA}) = \sum_{j=1}^N \sum_{i=1}^T \frac{k}{\exp(\omega_{max}^i - \omega_h^i)} \quad (3.5)$$

$\omega_{max}^i$  是特征堆栈中最大的权重， $\omega_h^i$  是无剪裁区域， $k$  是调整损失变化的剧烈程度的超参数，我们希望损失  $\mathcal{L}_{RA}$  更多地关注伪造部分。

对于最终分类层，团队使用了一个多层感知机，并使用了一个 cross-entropy 函数来作为约束，如下：

$$\mathcal{L}_{cls} = -(y \log(F) + (1 - y) \log(1 - F)) \quad (3.6)$$

我们设定了最终的总目标，即优化下列函数，通过训练三个编码器和分类器的参数，最小化下列式子的期望值。

$$\min_{\theta_{RA}, \theta_{cls}, \theta_{GR}} \omega \cdot \mathcal{L}_{RA}(\theta_{GR}, \theta_{RA}) + \mathcal{L}_{cls}(\theta_{cls}) \quad (3.7)$$

### (3) 基于轻量化的网络架构

用户在使用系统时，可能缺乏较高的算力支撑以及内存空间，我们需要对训练好的模型进行压缩，进行网络轻量化，在不改变网络结构的情况下，得到性能良好的目标模型，然后应用部署。我们提出如下的训练框架：

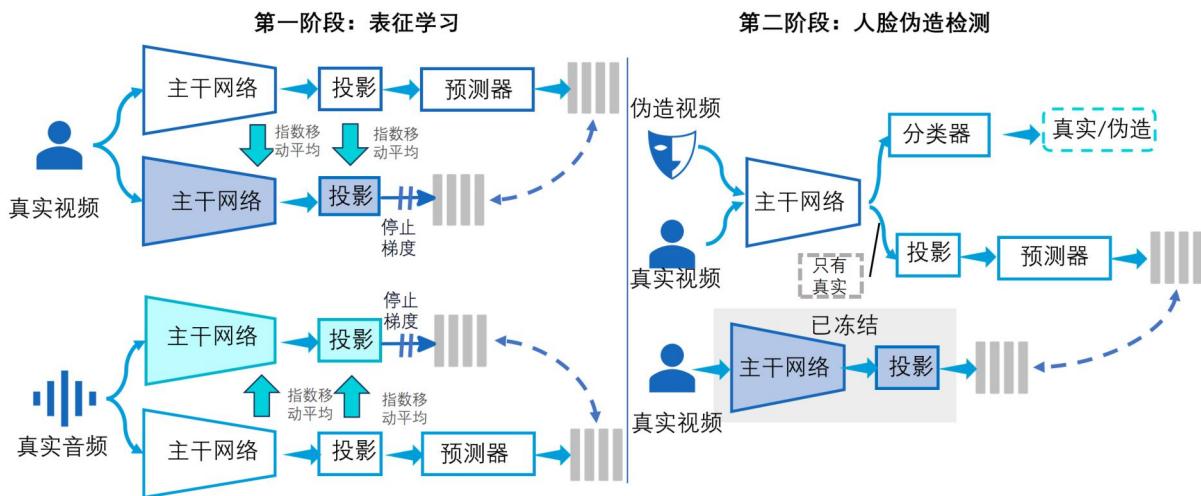


图 3.13 模型轻量化技术框架

在第一阶段，音频和视频的师生网络对通过自监督学习进行表征学习。到了第二阶段，音频教师模型被丢弃，视频教师模型生成目标供网络预测，同时网络进行伪造检测。

#### (a) 第一阶段：表征学习

当前阶段利用自我监督学习从大量自然的真实视频中学习说话人面部外观和行为信息等视频表征，这些表征随后在第二阶段中被用作预测目标，以规范二分类任务中的伪造检测。具体实现如下：

假设能够访问一个包含大量真实说话面孔的数据集  $\mathcal{D}_r$ 。一个样本  $x \in \mathcal{D}_r$  是一个视频  $x_v \in \mathbb{R}^{T_v \times H \times W \times 3}$  (其中  $T_v$  为视频帧数，高度为  $H$ ，宽度为  $W$ ) 及其对应的音频，以对数梅尔频谱图表示  $x_a \in \mathbb{R}^{T_a \times L}$  (其中  $T_a$  为音频帧数， $L$  为梅尔滤波器数量)，设置  $T_a = 4T_v$ 。

我们的架构包括音视觉两种模态的学生和教师模型对。教师模型生成的目标是其他模态的学生必须预测的。具体来说，视频和音频教师主干网络分别是  $f_v^t$  和  $f_a^t$ ，从输入中生成嵌入  $e_v^t = f_v^t(x_v)$  和  $e_a^t = f_a^t(x_a)$ ，然后通过投影函数  $g_v^t$  和  $g_a^t$ ，产生密集的视频和音频目标  $z_v^t = \text{norm}(g_v^t(e_v^t)) \in \mathbb{R}^{T_v \times C}$  和  $z_a^t = \text{norm}(g_a^t(e_a^t)) \in \mathbb{R}^{T_a \times C}$ ，其中  $C$  是嵌入的维度， $\text{norm}(\cdot)$  表示跨通道维度的  $l_2$  归一化。音频主干网络在时间维度上进行下采样，使视频和音频嵌入具有相同的形状。

学生模型的架构与对应的教师模型相同，但每个学生模型额外包含一个预测器，预测来自另一模态的目标。设视频和音频预测分别为  $p_v = \text{norm}(h_v(z_v^s))$  和  $p_a = \text{norm}(h_a(z_a^s))$ ，其中  $h_v$  和  $h_a$  表示预测器， $z_v^s$  和  $z_a^s$  是学生投影器后的未归一化表示。损失函数为：

$$\mathcal{L} = \frac{1}{2} \|\text{sg}(z_v^t) - p_a\|_F^2 + \frac{1}{2} \|\text{sg}(z_a^t) - p_v\|_F^2 \quad (3.8)$$

其中， $\|\cdot\|_F$  表示 Frobenius 范数， $\text{sg}$  表示“停止梯度”，强调目标被视为常量进行计算。总损失在所有样本上取平均，学生模型通过梯度下降进行优化，如果我们将视频教师模型的权重表示为  $\psi^v$ ，对应的学生模型权重表示为  $\theta^v$ ，那么在每次迭代中满足

$$\psi^v \leftarrow \mu\psi^v + (1 - \mu)\theta^v,$$

其中， $\mu$  是一个接近于 1 的动量参数，音频教师模型的权重更新类似上述过程。

### (b) 第二阶段：多任务伪造检测

在该阶段，教师模型在此阶段被冻结，使用第一阶段的视频教师模型生成的目标进行预测。我们在真实视频数据集  $\mathcal{D}_r$  的基础上，假设可以访问一个伪造视频数据集  $\mathcal{D}_f$ 。因此，我们的完整数据集为  $\mathcal{D} = \mathcal{D}_r \cup \mathcal{D}_f$ 。我们的架构由一个共享的骨干网络  $f$ （其权重为  $\theta_b$ ）和两个头部组成：一个用于伪造分类损失的监督头部（其权重为  $\theta_s$ ）和一个用于目标预测损失的辅助头部  $q$ （其权重为  $\theta_a$ ）。辅助损失表示为：

$$\mathcal{L}_a(\mathcal{D}_r; \theta_b, \theta_a) = \mathbb{E}_{x \sim \mathcal{D}_r} \|q(f(x^v; \theta_b); \theta_a) - t(x^v)\|_F^2, \quad (3.9)$$

其中， $t$  是第一阶段的教师，辅助头部和教师模型的输出如同第一阶段一样进行  $l_2$  正则化。监督损失  $\mathcal{L}_s(\mathcal{D}; \theta_b, \theta_s)$  是一个对数调整版本的二元交叉熵，以解决类不平衡问题。此外，为了获得对数，我们对特征向量和最后一层线性层的权重进行  $l_2$  正则化（并将其偏置设置为 0），获得一个余弦分类器。这更好地与辅助损失结合，后者也可以用余弦相似性来表示。最终的目标是：

$$\min_{\theta_b, \theta_s, \theta_a} \mathcal{L}_s(\mathcal{D}; \theta_b, \theta_s) + w\mathcal{L}_a(\mathcal{D}_r; \theta_b, \theta_a), \quad (3.10)$$

---

其中， $w$  是一个缩放因子，我们将其设置为 1。视频教师从第一阶段转移过来，并从此冻结。骨干网络的架构与第一阶段的视频骨干网络相同，我们用学到的权重初始化它。辅助头部包括一个随机初始化的投影器和预测器，如同第一阶段。监督头部是一个余弦分类器，如前所述。一个批次由 32 个伪造样本和 256 个真实样本组成，以有效利用更多的真实样本。我们使用 AdamP 优化器 [47]，学习率为  $3 \times 10^{-4}$ ，并采用与第一阶段相同的预处理和增强方法。我们训练了 150 个周期，并使用验证集进行早停。

最终，可以得到一个轻量化的基于时序不一致捕捉的伪造检测技术，不仅可以准确高效地识别 DeepFake 伪造内容，还可以解决 LipSync 伪造技术的难检测问题。

### 3.2.2 基于交叉模态情绪一致性的诈骗心理识别方法

当前，诈骗手段和方法日渐多样与复杂，诈骗分子会想尽一切诱导或伪造技术来绕过诈骗防御系统的检测。但无论如何，诈骗分子的诈骗动机或意图难免会表露在其动作、面部表情和语音情绪等多个情感表达方面。当下的一些情感识别技术往往只针对单一维度的情感细节，无法准确有效地体现真正的心理情绪，而且不同个体表达情绪的方式不同，单一维度的情感识别往往会遗漏情感细节，甚至在识别情感时完全失效。“一盾当关”系统在动作、面部表情、语音情绪等多个模态上综合交叉地分析检测对象的情感，并将捕捉到的异常情绪对应的语音文本进行对照分析，这样既能不遗漏情感细节，又能全面、整体性地分析出检测对象的诈骗心理与诈骗动机。

“一盾当关”系统的跨模态情绪识别技术通过动作捕捉、面部表情识别和语音情绪识别等多个角度综合判断情感。在面部表情识别方面，为了凸显通话视频的时间维度特征，防止系统将某一时间点的表情误判为一整段时间的情绪，同时实现高精度与高响应，我们将面部表情识别进一步分为动态面部情绪识别与静态面部情绪识别。各模块简介如下：

**①异常动作捕捉：**结合了 YOLO[48] 目标检测模型的空间定位优势和人体关键点检测的精细化分析能力，利用特征金字塔网络（FPN[49]）处理人体关键点信息，从而能够捕捉到检测对象的各种小动作。这些小动作可能包括手部的微小移动、头部的细微转动等，能够提供更多的情绪线索。

**②动态面部表情识别：**采用文本和图像双轨处理策略，特别考量图像帧之间的时间序列信息，以捕捉情绪表达的连续性和动态变化特征。通过分析视频中的面部表情变化，系统能够识别出情绪的微妙变化，例如从微笑到皱眉的过渡，最终输出视频载体上的情绪。

**③静态面部表情识别：**侧重于提升识别过程的响应速度和准确性，融合了 YOLO

和 Mamba[50] (一种基于状态空间的模型) 的优势, 采用最新的 FER-YOLO-VSS[51] (基于 YOLO 和视觉状态空间相结合的面部表情识别技术) 双分支模块。该模块能够在缺乏时间维度信息的情况下, 依然实现高效的情绪状态评估, 确保在静态图像中准确识别出面部表情。

**④语音情绪识别:** 通过深入分析语音信号的音调、节奏、强度等声学属性, 揭示语音中蕴含的情绪特征。系统能够识别出语音中的情绪变化, 例如愤怒、悲伤、喜悦等, 补充了非言语渠道的情绪信息, 使情绪识别更加全面。

最后, 我们综合分析动作捕捉、动态或静态面部表情、语音情绪的判定结果, 定位通话过程中异常情绪的时间点, 并结合相应的语音文本诈骗分析, 得到检测对象的诈骗心理与诈骗动机。通过这种多模态情绪识别技术, 系统能够更准确地识别出潜在的诈骗行为, 提高防御诈骗的效果。

以下是我们详细的说明。

### (1) 动作捕捉

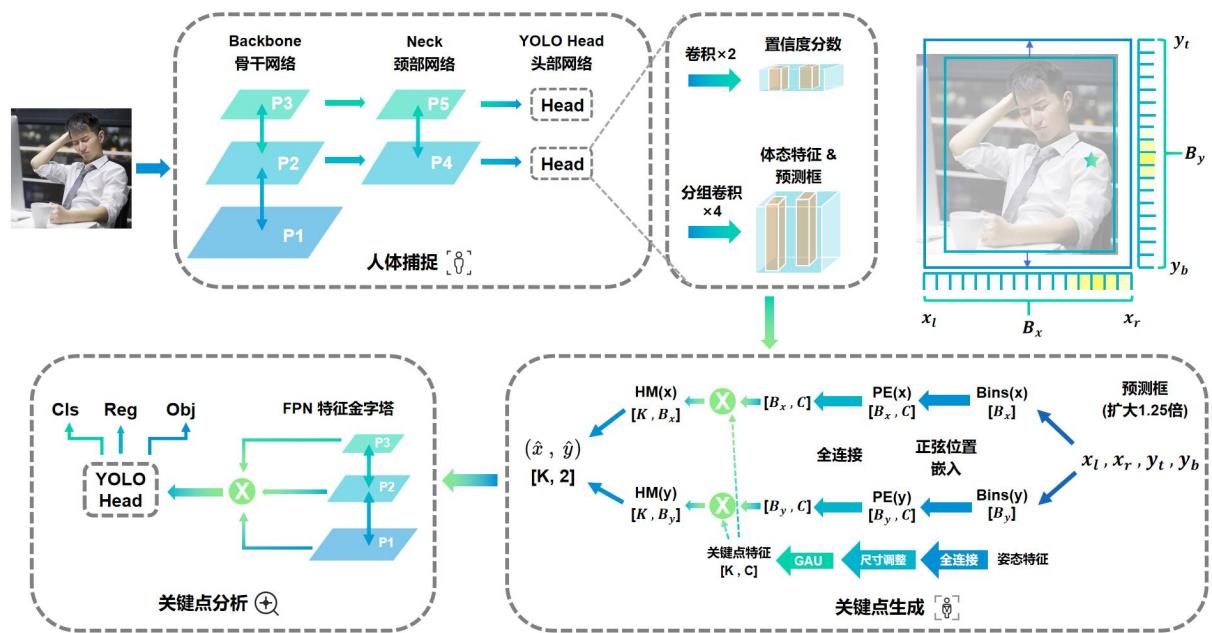


图 3.14 模型整体框架

动作捕捉模块分为人体捕捉、关键点生成和关键点分析三个部分, 各个部分的作用见表3.1:

---

表 3.1 组成部分及其作用

组成部分	作用
人体捕捉	提取输入图像特征，框选出图像中包含人的部分、置信度分数还有姿态特征。
关键点生成	根据标注框生成关键点坐标热图，据此提取出每个关键点的坐标，并且根据关键点的可见性过滤或优化被遮挡的关键点（考虑到遮挡和拥挤的场景）
关键点分析	将关键点位置送入特征金字塔中由粗到细的逐层提取出检测对象的身体动作信息，并结合每一层的输出综合提取检测对象的动作特征，并送入 YOLO Head 进行动作的分类。

在**人体捕捉**部分，图像特征的提取和处理是通过精心设计的三个主要部分来实现的：Backbone（骨干网络）[52]，用来从输入图像提取多尺度特征图；Neck（颈部网络）[49]，用来融合不同的多尺度特征图来生成丰富的多尺度特征表示；Head（头部网络）[48]，处理融合后的特征图，预测置信度分数并精细化提取人体姿态特征。

### (a) Backbone (骨干网络)

本技术的骨干网络采用 CSPDarknet 算法 [53]，负责从输入图像中提取特征。这些特征图根据尺度大小被划分为 P1、P2、P3 三层，其中尺度越大，特征图的空间分辨率越高。这种设计有效减少了特征提取过程中的计算量，同时保持了对空间细节的高分辨率表达，为后续的特征融合和提取奠定了基础。

### (b) Neck (颈部网络)

在颈部网络中，我们采用 Hybrid Encoder 混合编码器 [54] 来融合和提取特征图。这一过程通过整合来自 Backbone 的三层不同尺度的输出，生成更为丰富和全面的多尺度特征表示。Neck 层的输出是进一步融合后的 P4 和 P5 两层特征图，其中 P4 具有更大的尺度，而 P5 则提供更精细的特征信息，这样的设计在降低计算成本的同时，确保了信息的丰富性和细节的精确性。

### (c) Head (头部网络)

头部网络利用 Neck 层提供的融合特征图进行深入处理。首先，通过两次卷积操作来预测置信度分数，这些分数反映了预测框对应目标存在的概率。随后，通过四次分组

---

卷积来精细化提取人体姿态特征，并据此框选出人体的关键部位。分组卷积的策略允许模型在不同的通道组中独立学习特征表示，这不仅减少了计算复杂度，还增强了模型对特征多样性的捕捉能力。

在关键点生成部分的核心是动态坐标分类器 (Dynamic Coordinate Classifier, DCC) [55]。DCC 首先进行动态坐标箱分配 (Dynamic Bin Allocation) [56]，将图像上的特定区域划分为一系列离散的坐标箱 (bins)。这些坐标箱用于表示关键点可能的位置，并与关键点出现在这些位置的概率相关联。初始的边界框通过逐点卷积层进行回归预测，并扩展以确保覆盖所有关键点，即使在预测存在误差的情况下也能有效覆盖。这些扩展后的边界框在水平和垂直方向上均匀划分为  $B_x$  和  $B_y$  个坐标箱，每个坐标箱的 x 坐标计算公式如下：

$$x_i = x_l + (x_r - x_l) \frac{i - 1}{B_x - 1} \quad (3.11)$$

其中  $x_r, x_l$  分别是边界框的左右两侧，索引  $i$  从 1 变化到  $B_x$ 。y 轴上的箱计算方式类似，公式如下：

$$y_i = y_l + (y_r - y_l) \frac{i - 1}{B_y - 1} \quad (3.12)$$

分配好坐标箱后就会进行坐标箱的动态编码 (Dynamic Bin Encoding)，作用是根据人体实例的大小和形状自适应地调整坐标箱的表示。动态编码公式如下：

$$[PE(x_i)]_c = \begin{cases} \sin\left(\frac{x_i}{t^c/C}\right), & \text{for even } c \\ \cos\left(\frac{x_i}{t^{(c-1)/C}}\right), & \text{for odd } c, \end{cases} \quad (3.13)$$

在 DCC 的框架内，我们引入了一个温度参数  $t$  来调整坐标箱编码的敏感度。这里， $c$  表示坐标箱的索引，而  $C$  是坐标箱在每个维度上的总数。对于垂直方向的坐标箱，计算方法与水平方向保持一致，公式如下：

$$[PE(y_i)]_c = \begin{cases} \sin\left(\frac{y_i}{t^c/C}\right), & \text{for even } c \\ \cos\left(\frac{y_i}{t^{(c-1)/C}}\right), & \text{for odd } c, \end{cases} \quad (3.14)$$

下一步，为了进一步提升位置编码的精细度和适应性，我们采用了全连接层对编码进行细化，通过可学习的线性变换  $\phi$ ，提高编码在 DCC 中的有效性，使模型能够更加精准地预测关键点的位置。

---

经过全连接层的处理，每个关键点的特征向量  $f_k$  被进一步映射并与坐标箱编码相结合，生成一个 1-D 的热图分布。这个热图分布反映了关键点在各个坐标箱中的概率分布。最后通过 *softmax* 函数将这个概率分布转换为归一化的置信度，确保所有坐标箱的概率和为 1。

$$\hat{p}_k(x_i) = \frac{e^{f_k \cdot \phi(PE(x_i))}}{\sum_{j=1}^{B_x} e^{f_k \cdot \phi(PE(x_j))}}, \quad (3.15)$$

其中  $f_k$  是第  $k$  个关键点的特征向量。关键点特征的提取是通过头部网络进行的，该网络不仅提取姿态特征，而且通过一次全连接层和尺寸调整，进一步精细化这些特征。随后，这些特征被输入到门控注意力单元（Gated Attention Unit, GAU）[57]，GAU 利用注意力机制集中于输入特征中最关键的部分，强化了关键特征并抑制了无关的噪声，从而显著提升了特征的表达力。

在热图的扫描过程中，我们通过寻找  $B_x$  和  $B_y$  两个方向上坐标箱热图中的峰值点，精确地确定每个关键点的位置。然而，为了应对遮挡和拥挤等复杂情况，我们在 YOLO 模型提取的单元格信息基础上，进一步分析每个关键点的可见性，优化或过滤掉不可见的关键点，确保了关键点坐标信息的准确性。

在关键点分析阶段，我们采用了特征金字塔网络 [49] 来提取关键点信息的特征。通过三层不同尺度的处理，由粗到细地分析关键点，从整体骨架到局部的脸部和手部细节，这种方法能够更全面和精确地提取动作特征。之后将提取到的动作特征输入到 YOLO Head 中，进行最后的动作分类与目标框标注。

最终，通过这一系列精细调整和优化的过程，我们的动作捕捉技术成功地输出了动作信息，完成了模型的整个任务。我们团队的动作捕捉技术，结合了 YOLO 模型的快速响应和轻量化优势，人体关键点检测的精确性和细致性，以及特征金字塔的关键点信息整合能力，能够在减小计算量的同时，实现非常高的识别准确率。它能够在各种特殊场合下准确预测包括摸脖子、挠头、摸鼻子和斜视在内的细微动作，完全符合我们对动作捕捉技术的预期目标。

## (2) 动态面部情绪识别

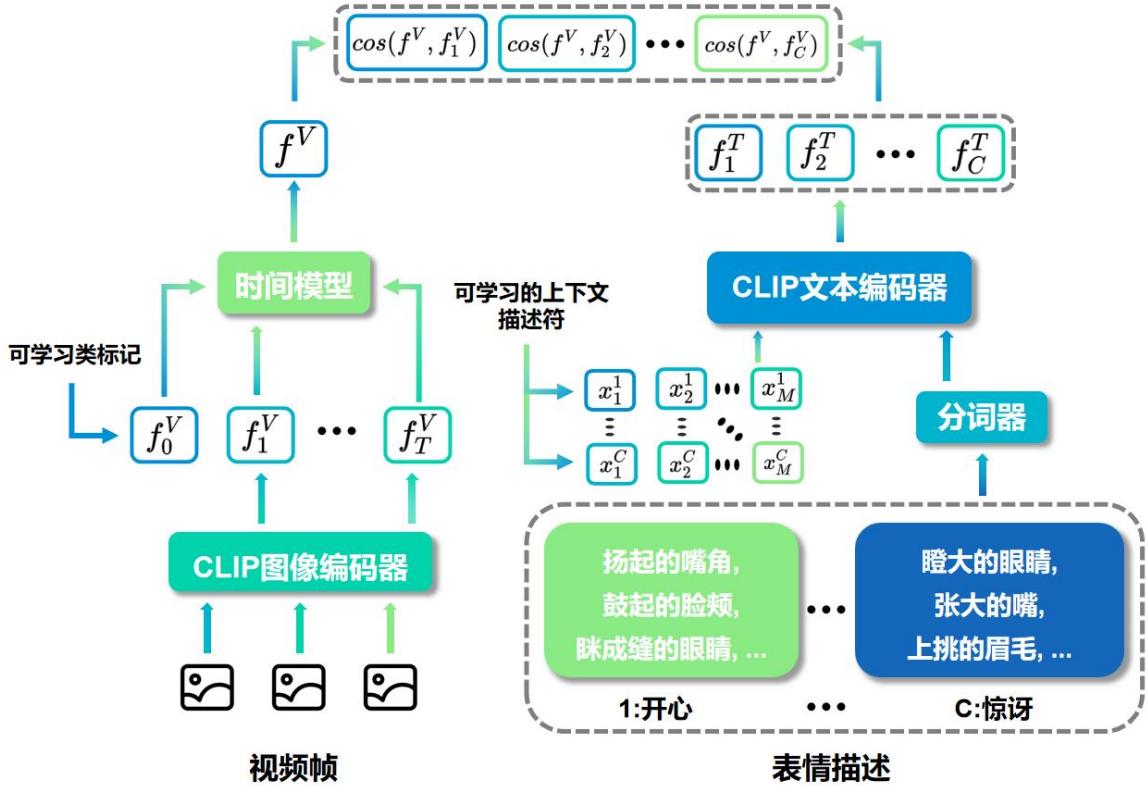


图 3.15 模型整体框架

如上图所示，我们将识别过程分为视觉部分和文本部分两个部分进行处理。两个部分的作用如表3.2:

表 3.2 模型构成及其作用

模型构成	作用
视觉部分	利用 CLIP 编码器捕捉帧级面部特征，然后再使用由多个 ViT 编码组成的时间模型对捕捉到的面部特征进行建模，最后得到最终的视觉表示 $f^V$
文本部分	将面部表情的动作描述作为文本编码器的输入，结合提取出的人脸特征进行分析，使预测更加全面完整（因为某些情绪特征都有相似的特征，比如快乐和惊讶的表情都有扬眉的动作，所以应当抓住每个情绪的特有特点）

在视觉部分，我们采用了 CLIP 视觉编码器来提取视频中每一帧的特征。这些帧级特征，连同额外的可学习类标记，将共同输入到我们设计的时间模型中。时间模型由多个 ViT 编码器构成，专门用于捕捉视频帧之间的时间关联和动态变化。

为了进一步丰富时间模型的能力，我们在模型中引入了可学习的位置嵌入，这使得模型能够编码每一帧在视频序列中的时间位置信息。通过这种方式，我们的模型不仅能够理解每一帧的内容，还能够把握帧与帧之间的时间顺序和上下文关系，最终生成综合的视觉表示。

具体实施过程中，我们从视频中采样  $T$  帧，得到输入数据  $x \in \mathbb{R}^{T \times 3 \times H \times W}$ 。对于视频中的每一帧  $x_i$ ，我们首先使用 CLIP 图像编码器  $f(\cdot)$  提取其特征向量  $f_i^v \in \mathbb{R}^L$ ，其中  $i \in \{1, 2, \dots, T\}$ ， $L$ ， $L$  代表特征向量的长度。随后，这  $T$  个特征向量将被送入时间模型  $g(\cdot)$  进行深入的时间序列建模，得到了最终的视觉表示  $f^V \in \mathbb{R}^L$ ：

$$f_i^v = f(x_i) \quad (3.16)$$

$$f^V = g(f_0^v + e_0, f_1^v + e_1, \dots, f_T^v + e_T) \quad (3.17)$$

其中  $f_0^v$  是类 *token* 的特殊可学习向量， $e$  表示为编码时间位置而添加的可学习位置嵌入。

在文本部分，我们利用与面部行为相关的描述，如表3.3：

**表 3.3 表情与描述**

表情	描述
快乐 😊	微笑的嘴唇，抬起的脸颊，皱起的眼睛和弯曲的眉毛。
悲伤 😞	眼泪，下垂的嘴巴，下垂的上眼睑和皱起的额头。
平静 😌	放松的面部肌肉，平直的嘴唇，光滑的额头和不显眼的眉毛。
愤怒 😡	皱起的眉毛，狭窄的眼睛，紧闭的嘴唇和张开的鼻孔。
惊讶 😱	睁大的眼睛，张开的嘴巴，抬起的眉毛和僵住的表情。
厌恶 😤	皱起的鼻子，低垂的眉毛，紧闭的嘴唇和狭窄的眼睛。
恐惧 😰	抬起的眉毛，张开的嘴唇，皱起的眉头和缩回的下巴。

此外，我们还采用了可学习的提示作为每个类的描述符的上下文。提示符的形式如下：

$$P_K = [p]_k^1 [p]_k^2 \cdots [p]_k^M [Tokenizer(description)]_k \quad (3.18)$$

其中  $M$  是一个指定上下文 token 数量的超参数,  $k \in \{1, 2, \dots, C\}$ ,  $C$  是面部表情类的数量, 每个  $[p]_k^m$ ,  $m \in \{1, 2, \dots, M\}$ , 是一个与词嵌入相同维度的向量。这里, 我们采用类特定的上下文, 其中上下文向量独立于每个描述。通过将提示  $P_k$  转发给文本编码器  $h(\cdot)$ , 我们可以得到表示视觉概念的  $C$  分类权重向量  $f_k^T \in \mathbb{R}^L$ , 公式如下:

$$f_k^T = h(P_k) \quad (3.19)$$

那么预测概率可以计算为:

$$p(y = k|x) = \frac{\exp(\cos(f^V, f_k^T)/\tau)}{\sum_{k'=1}^C \exp(\cos(f^V, f_{k'}^T)/\tau)} \quad (3.20)$$

其中  $\tau$  为 CLIP 学习到的温度参数,  $\cos(\cdot, \cdot)$  为余弦相似度。就这样, 我们的模型结合了图像的帧特征和语言描述完成了情绪的预测。我们的模型核心优势在于其卓越的时间关系处理能力, 以及能够基于每个表情的特征描述进行面部表情识别的能力。这种结合了时间动态和表情特征描述的方法, 赋予了模型在面对光线极端或面部遮挡等特殊条件时的显著鲁棒性和泛化能力。

### (3) 静态面部表情识别

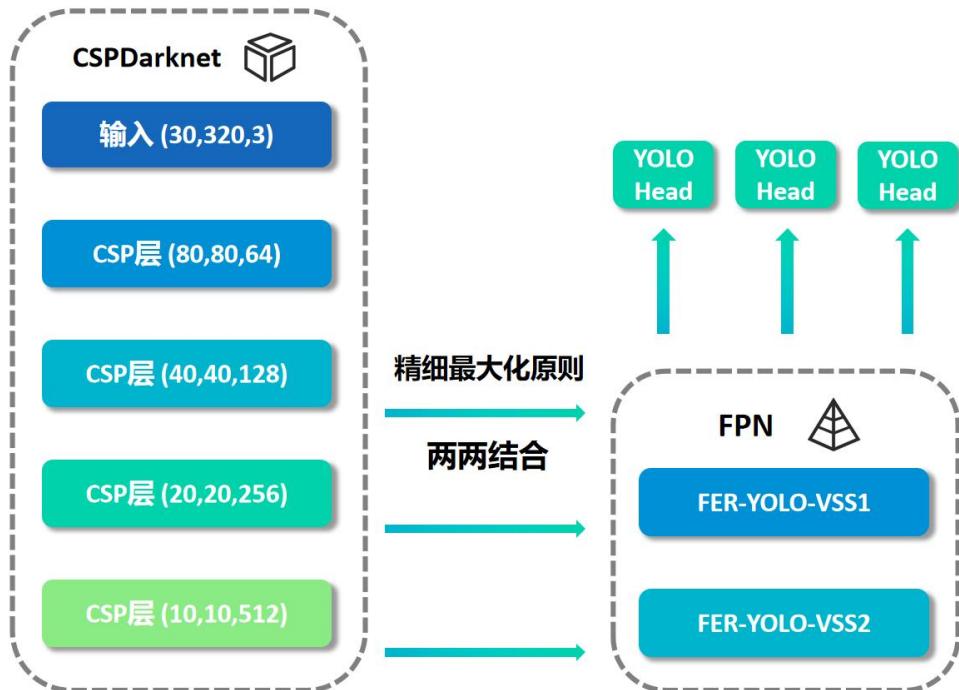


图 3.16 静态面部情绪识别模型整体架构

---

我们的静态面部情绪识别模型包含 CSPDarknet、FRN 和 YOLO Head 三个核心模块，每个模块的功能如下表：

表 3.4 模型构成及作用

模块名称	作用
CSPDarknet	将输入图像转换为三个不同尺度的特征图，输出从粗糙到精细的层次多级特征
FPN	有效地融合 CSPDarknet 中的跨尺度特征，捕获不同层次的细节和上下文信息，从而增强整体特征表示
YOLO Head	承担着分类和定位的双重责任，单独分析 FPN 生成的多尺度特征图上的特征点，确定它们与目标对象的关联

**CSPDarknet[58]** 相对其他的 CNN 卷积神经网络，采用了 CSPNet(Cross Stage Partial Network，跨阶级部分连接)的方式来减少特征提取的计算量并提高效率。CSP 结构将网络的层分为多个路径，每个路径包含一系列卷积层，这些层独立地处理输入特征图。这些路径主要由“small”和“big”路径组成，“small”路径包含  $1 \times 1$  卷积操作用于降维，减少特征图的通道数，从而减少后续层的计算负担；“big”路径包含  $3 \times 3$  卷积操作，能够在保持特征图深度或通道数的同时，提取更高级的特征。

CSP 模块通过将“small”和“big”路径的输出融合来实现特征融合，若令本模块的输入为  $x$ ，则合并的操作可表示为：

$$y = f_{small}(x) + f_{big}(x) \quad (3.21)$$

其中  $f_{small}(x)$  表示经过“small”路径处理后的特征图， $f_{big}(x)$  表示经过“big”路径处理后的特征图， $y$  是合并后的特征图，能够结合两条路径的信息。

在我们的网络中，决定提取三个 CSP 层的输出来进一步的提取特征，在每一层前都需要通过二维卷积、批量归一化、SiLU 激活来处理该层的输入，从而使网络能够学习和模拟复杂的函数映射。若令  $x_0$  表示最初的输入， $x_0$  大小为  $320 \times 320 \times 3$ ， $F_k(\cdot)$ ， $k \in \{1, 2, 3, 4\}$  表示每一层的处理， $x_k = F_k(x_{k-1})$ ， $k \in \{1, 2, 3, 4\}$  表示每一层的输出，其中  $x_1$  大小为  $80 \times 80 \times 64$ ， $x_2$  大小为  $40 \times 40 \times 128$ ， $x_3$  大小为  $20 \times 20 \times 256$ ， $x_4$  大小为  $10 \times 10 \times 512$ ，我们取  $x_2, x_3, x_4$  三层输出作为 FPN 输入。从  $x_2$  到  $x_4$ ，它们的特征表现为从粗糙到精细。

前文已经介绍过 FPN，这里我们依然使用 FPN 进行处理。在 FPN 中，会将输入  $x_2, x_3, x_4$  通过上采样或者二维卷积来适应彼此的尺寸，并按照“精细最大化”的原则，将他们两两沿着通道维度连接并送入 FER-YOLO-VSS 来提取面部特征。若用  $g(\cdot, \cdot)$  表示将两个输入适应尺寸后相连接并送入 FER-YOLO-VSS1[51] 进行处理， $h(\cdot, \cdot)$  表示连接并送入 FER-YOLO-VSS2[51] 进行处理。则处理的顺序可以表示为：

$$\begin{aligned} y_1 &= g(x_3, x_4) \\ y_2 &= g(x_2, y_1) \\ y_3 &= h(y_1, y_2) \\ y_4 &= h(x_4, y_3) \end{aligned} \quad (3.22)$$

其中  $y_2, y_3, y_4$  会被送入 YOLO head 来进行最终的预测，他们的大小分别为  $40 \times 40 \times 128, 20 \times 20 \times 256, 10 \times 10 \times 512$ 。FPN 的核心就在于使用 FER-YOLO-VSS(Facial Expression Recognition - YOLO Visual State Space) 模块进行面部表情检测和分类，该模块是一个双分支结构，输入首先通过通道分割处理，分成两个等大小的子输入，然后并行地输入特征细化模块 (FRM)[59] 分支和全向状态空间 (OSS)[60] 分支进行处理，我们就是在 OSS 分支中使用 Mamba 模型进行状态空间建模，来处理长距离依赖问题的。这种处理方式能够更有效地捕捉和提取图像中的关键特征信息。

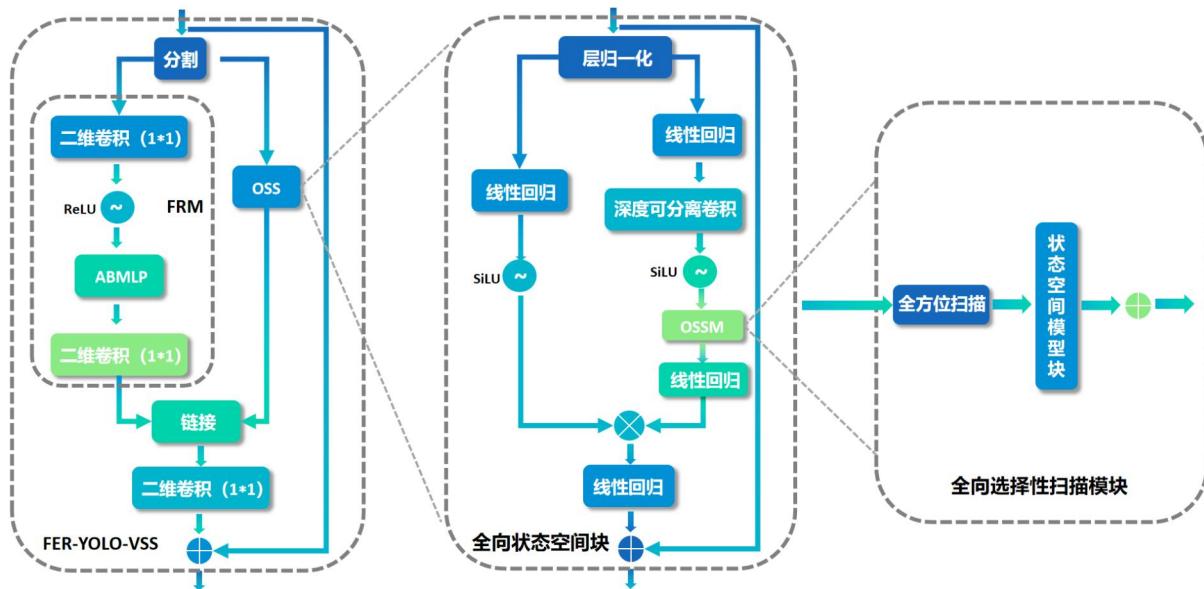


图 3.17 FER-YOLO-VSS 模型

FRM 分支采用了连续的通道维度压缩策略，用以增强模型学习区分性和上下文感知的特征表示的能力。此外，该分支还整合了一个注意力机制 (ABMLP, Attention Block

with Multi-Layer Perceptron)[61]，通过自适应特征权重调整来调整不同特征的重要性。经过这一系列的处理后，FRM 分支最终恢复到原始的通道数，从而确保信息的完整性和准确性。

OSS 分支首先对输入进行层归一化处理，后将处理的特征又分为两个并行的子路径。在第一个路径中，特征只经历线性变换层和激活函数两个简单的转换。第二个路径会相对复杂，特征会经历线性层、深度可分离卷积和激活函数三个级别的逐步处理，然后进入全向选择性扫描模块（OSSM）[62] 以深入提取特征信息。OSSM 利用状态空间模型（State Space Model，SSM，也就是我们使用的 Mamba 模型）技术在水平、垂直、对角线和反向对角线方向上对面部表情图像进行双向选择性扫描。八种不同方向的选择性扫描能够从多个方向捕获大规模空间特征。在此之后，应用层归一化来标准化特征，并且这个分支的输出通过逐元素乘法与第一个分支的输出深度融合。随后，通过线性混合层整合每个分支的特征，并采用残差连接策略，得到了 OSS 模块的输出。

最后两个分支的输出特征沿着通道维度进行拼接，并通过  $1 \times 1$  卷积层进行深度特征融合，来增强特征图之间的深层交互效应，得到了最终的输出。FER-YOLO-VSS1 和 FER-YOLO-VSS2 是 FER-YOLO-VSS 的两个变体，FER-YOLO-VSS1 旨在减少通道数，从而减少计算的复杂度；FER-YOLO-VSS2 会保持输入和输出的通道数一致，更注重于特征的细化和上下文信息的整合。总之，FER-YOLO-VSS 旨在实现局部和全局信息的互补融合，通过注意力机制增强模型处理关键信息的能力，从而提高整体性能。



图 3.18 YOLO Head 结构

YOLO Head 是对最终的图像特征进行预测、分类与定位，我们将 YOLO Head 分为

两条路进行处理，每条路都有多个卷积层 (Conv2D)、批量归一化 (BN) 和 SiLU 激活函数组成，用来提取高级特征。第一条路径 (Cls 分支) 用来输出面部表情的置信度向量，也就是在每个类别中的概率；第二条路径中的 Reg 分支用来预测目标的位置，也就是边界框的坐标，Obj 分支用来预测对象存在的概率。以此得到模型最终的输出。

FER-YOLO-Mamba[51] 模型通过结合 YOLO 目标检测技术的快速和精确特性以及 Mamba 模型在状态空间建模中的优势，能够快速、准确的识别图片中诈骗对象的面部表情，完美的迎合了我们团队对静态面部表情识别的需求。

#### (4) 语音情绪识别技术

语音情绪识别本质上是一个多分类任务，难点在于不同语言之间的迁移效果和不同群体之间的推理性能。不同年龄群体之间的语音特征差异非常明显，跨群体推理需要先进行跨群体数据增强。之前有关语音情绪识别的研究大部分集中在英语或者欧洲其他语言上，利用深度神经网络识别中文语音情绪的工作相对少见。我们团队设计一种基于跨群体数据增强的中文语音情绪识别技术，探索音频特征在深度神经网络中的有效性。

首先对音频数据进行特征提取，我们考虑如下音频特征：

##### (a) 频谱质心

频谱质心是描述音色特征的重要参数，用在频谱图中的频谱质心，可以表达各帧的频率分布趋势。将各帧的频谱质心绘制成波形图，如图3.19。

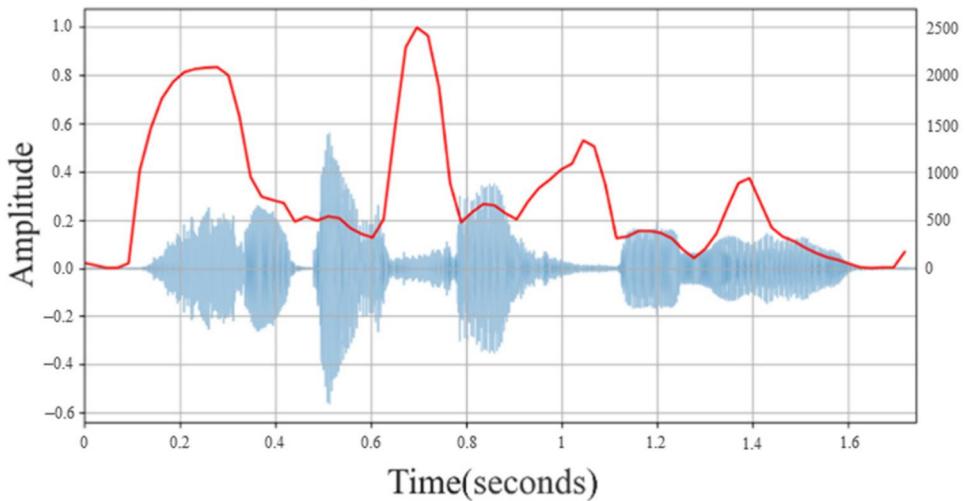


图 3.19 样本数据的频谱质心

从物理意义上讲，频谱质心可以描述声音的明亮程度，当声音浑厚深沉时，频率偏向低频，频谱质心就比较低；当声音明亮轻快时，通常集中在高频，频谱质心就比较高。

##### (b) 频谱平坦度

频谱平坦度表示音频频带间能量分布的平均程度。将频谱分为 N 个频带，其中  $x(n)$  表示第 n 个频带的总能量强度，然后计算  $x(n)$  的几何平均值和算术平均值，并将变化率表示为比值，公式如下：

$$SF = \frac{\sqrt[N]{\prod_{n=1}^N x(n)}}{\frac{\sum_{n=1}^N x(n)}{N}} \quad (3.23)$$

由于算术平均值大于几何平均值，因此计算结果在 0 ~ 1 之间。当各个频带的能量分布比较平均时，比值会趋近于 1，反之则趋近于 0。

### (c) 频谱对比度

将频谱图中的每一帧划分为多个子带，通过计算子带内频谱峰和频谱谷的平均能量（即峰值能量和谷值能量）得到能量对比度，即频谱对比度。对比度高表示声音信号清晰、带状信号较窄，对比度低表示有噪声。

### (d) 色度特征

色度特征是一个统称，指的是色度向量和色度图。色度向量包含 12 个元素，分别为 \*C\*、\*C#\*、\*D\*、\*D#\*、\*E\*、\*F\*、\*F#\*、\*G\*、\*G#\*、\*A\*、\*A#\* 和 \*B\*。\* 这些元素表示一个周期（比如一帧）内 12 个声音级别的能量。将不同八度的同一声音级别的能量累积起来，色度图就是色度向量序列。色度向量由 12 个元素特征的向量组成，用来表示信号中每个音阶的能量。可视化的色度图如图3.20：

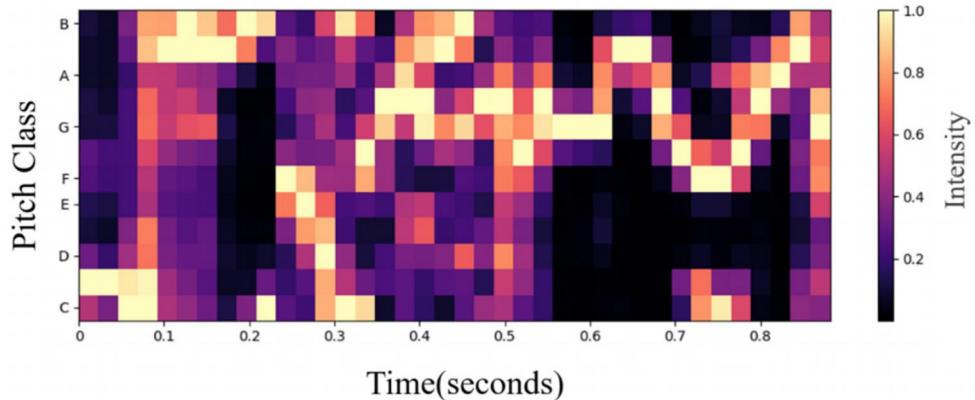


图 3.20 样本数据的色度特征

(e) 过零率。过零率（Zero-Crossing Rate）是指每帧中音频经过零点的次数。其公式如下：

$$ZCR = \frac{1}{T-1} \sum_{t=1}^{T-1} \pi \{ s_t s_{t-1} < 0 \} \quad (3.24)$$

---

为了增强模型的跨群体推理能力，我们使用了跨群体数据增强，首先定义语言组为  $L = \{l_1, l_2, \dots, l_n\}$ ，年龄组为  $A = \{a_1, a_2, \dots, a_m\}$ ，对于语言为  $L_i$ ，年龄为  $a_j$  的群体，训练数据样本集和评估数据样本集分别为  $X_{l_i, a_j}^{trn}$  和  $X_{l_i, a_j}^{tst}$ 。跨群体数据增强技术，包括旋转、调谐、裁剪、加噪。这些转换不会改变原始数据的标签，但会增加原始类别中数据的可变性。具体来说，旋转操作是以 10% 为单位，将频谱前后段的音频进行顺序调换。调谐，是在保持整体时间和频率不变的条件下，改变数据的音调。裁剪是指对原始语音信号进行预加权，我们先采用一组三角带通滤波器获取对数能量，然后通过余弦转换得到不同频段的代表系数。模型训练采用交叉熵损失函数，优化器采用 AdamW，权重衰减率为 0.2，我们设置学习率为 0.0001，批处理大小为 36。此外，我们为了避免过拟合，增加了 dropout 操作。

### 3.2.3 基于文本诱导性特征捕捉的可解释性检测技术

常见的实时或离线交互场景下的诈骗依然采用基于事实引导的传统诱骗模式，但随着各种新颖话术的出现，尤其是定制化诈骗的出现，利用其强诱导性、隐蔽性和交互型，逐步获取受试者的信任，这使得受害者渐渐放下戒备而落入诈骗陷阱且不易察觉。我们深入分析了各种“基于事实”的引导式诈骗的发展过程，发现其核心本质，如“诱导转账”和“冒充”等，始终都未改变，这使得诈骗话术尽管多样化，但仍呈现出一定的模板性和语义上的高度相关性。

基于此，我们调研了目前的诈骗文本检测的大语言模型技术，发现并不存在一个专注于抓取诈骗本质特征的专业向模型。现有的模型都只能基于大规模数据集的训练，依赖于预定义的模式和特征，模型准确率低、可解释性不强，无法捕捉到诈骗话术的统一特征和固定化模板。一旦诈骗话术稍微改变，就无法保证检测的准确性和及时性，且再度训练会耗费大量资源。

因此，我们旨在提出一种专注于诈骗文本的本质特征的检测手段，通过“特征分类-内容分析-风险解释”的方式，以达到面对各种复杂诈骗和新型诈骗话术时，都能通过抓住其话术特征来提示诈骗风险，并提供内容针对性的具体此类特征诈骗的解释。诈骗话术识别过程如下图3.21：

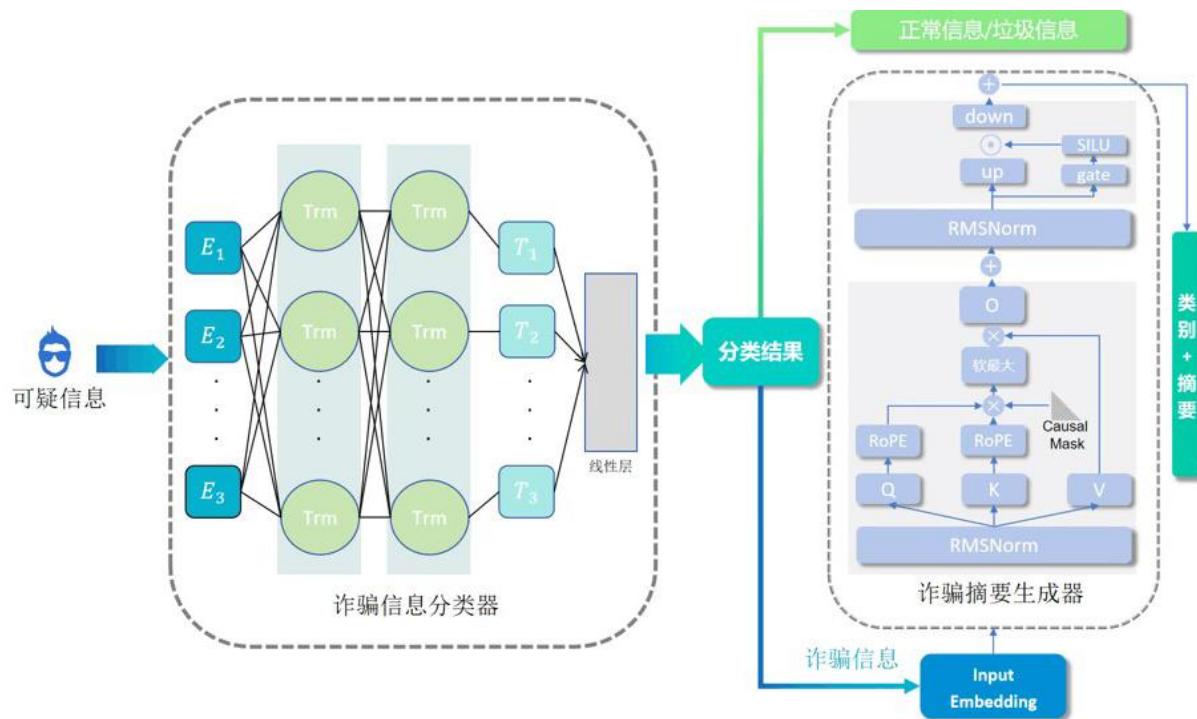


图 3.21 诈骗文本识别过程

我们知道，模型的对输入文本的理解分类能力是进行分析的基础，而后的摘要生成又依赖目前的大语言模型。因此，我们设计分类器技术架构以能增强对长输入文本的理解能力，可以对诈骗文本进行准确的分类工作；并在现有大语言模型上，修改模型架构，通过下一步的微调训练使模型可以通过抓住诈骗本质来理解文本并生成摘要内容。为保证模型后续的高效实用，我们也进行模型量化工作。

以下是技术实现细节。

### (1) 诈骗文本分类器

诈骗文本作为输入，通过嵌入层转换、编码器编码、池化层表示，得到对应输出。过程如下图3.22：

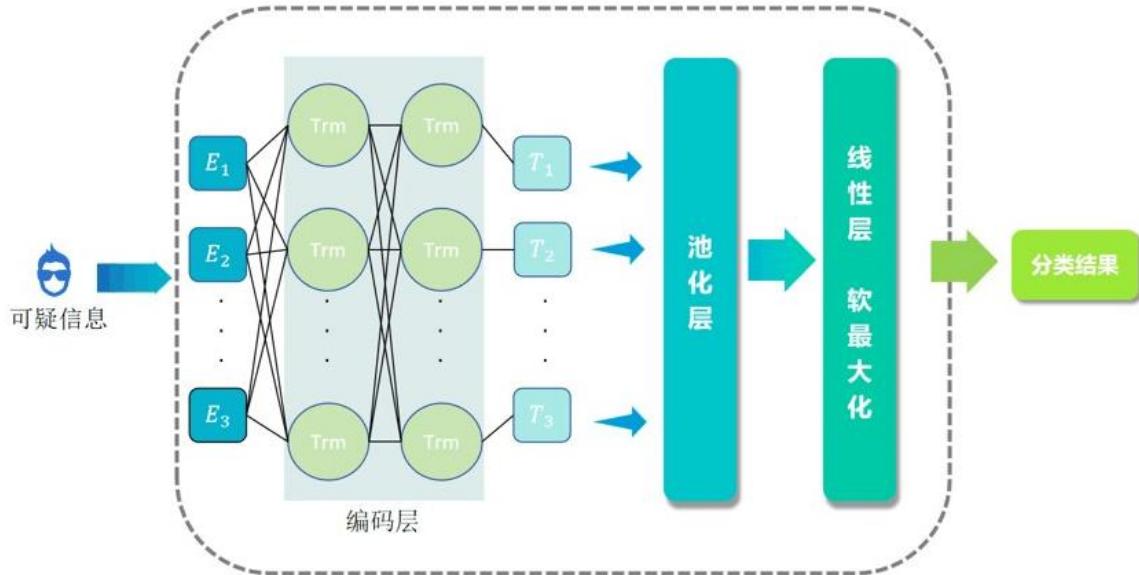


图 3.22 文本分类过程

### (a) 模型输入

受 BERT 的研究启发，在对文本数据预处理后，我们设置三个嵌入层，分别为词片嵌入层 (WordPiece Embedding)、位置嵌入层 (Position Embedding)、分段嵌入层 (Segment Embedding)。通过三个嵌入层的处理后，每个输入序列将会获得三种不同的向量表示，最终按元素相加，得到最终向量形式，并输入到编码层中。整体过程如图3.23：

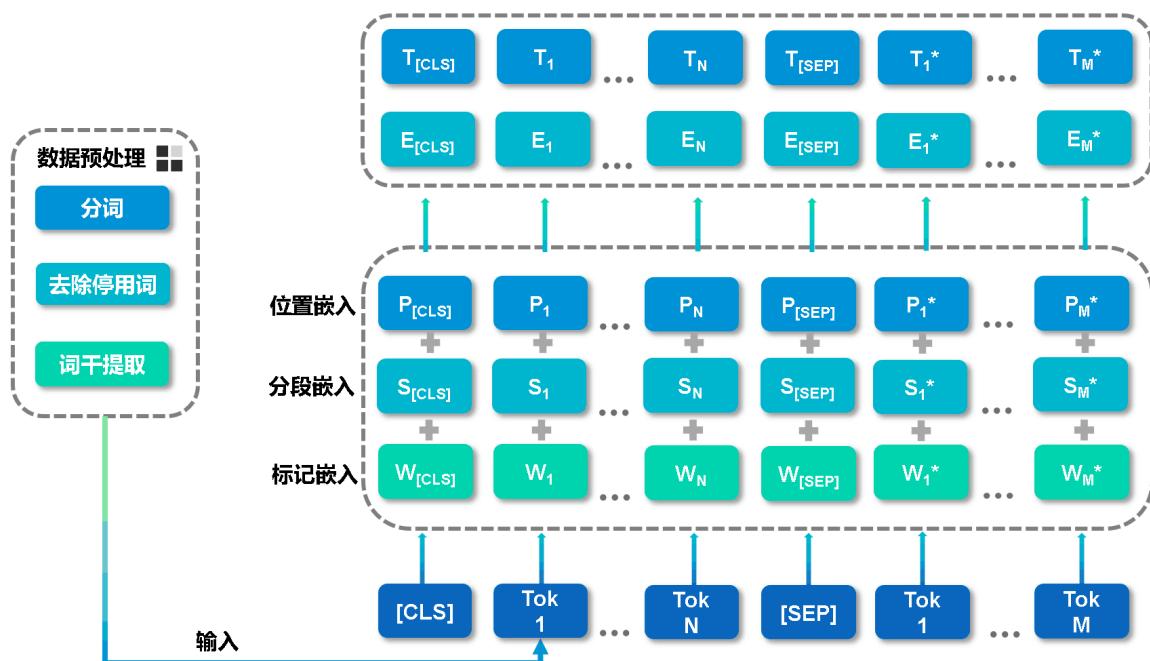


图 3.23 文本输入处理流程图

---

词片嵌入层，将单词分解成子词单元，并将这些子词映射到一个高维向量空间。通过子词分解和嵌入，提供细粒度的语义信息，确保即使遇到罕见或未知的词，模型也能通过其组成部分进行有效处理，以此来增强了模型的灵活性和泛化能力。

位置嵌入层，编码单词在序列中的位置信息，引入一个可训练的位置嵌入矩阵，使模型能够理解词语在句子中的顺序。因为并行处理输入的缘故，Transformer 模型本质上是无序的，而通过添加位置信息、引入序列，使模型能够捕捉词语的顺序，从而理解上下文关系和语义结构。

分段嵌入层，可以区分输入文本的不同部分，特别是在处理多个句子或文本片段时具有显著的作用，帮助模型理解和处理由不同来源或不同部分组成的输入。通过该嵌入层，我们使模型可以应对句子级任务和应答过程。

我们将单词转换维度也固定为 768 维度，以此在模型的性能和计算复杂度之间取得平衡，即不仅能够提供足够的表示能力来捕捉单词和子词的语义信息，能有效区分不同的词，还不会明显增加计算成本。经过三个嵌入层的转换后，我们会得到三种不同的向量表示，即：

- 词片嵌入， $(1, n, 768)$ ，词的向量表示
- 位置嵌入， $(1, n, 768)$ ，辅助 BERT 区别句子对中的两个句子的向量表示
- 分段嵌入， $(1, n, 768)$ ，让 BERT 学习到输入的顺序属性

我们再将三个向量依照元素位置对应关系对照相加，得到一个大小为  $(1, n, 768)$  的合成表示，以此作为编码层的输入，使模型在处理多样化和复杂的自然语言任务时具有更强的灵活性和泛化能力。

### (b) 模型输出

我们采用以下类型的模型输出：

词源表示：每个输入标记经过多头自注意力（Multi-Head Attention）和前馈神经网络（Feed-Forward Neural Network）层后，都会生成一个输出表示。这些表示是高维向量，捕捉了标记本身的语义信息以及与序列中其他标记的关系。对于序列中的每个标记，模型都会输出一个对应的表示。

序列表示：整体序列的表示通常指的是经过模型处理后，能够代表整个输入序列信息的向量。这种表示可以用于捕捉整个序列的语义信息，常用于分类任务。

最后，为了完成分类任务，我们在模型最后加一个线性层。具体操作为先获取 [CLS] 标记的表示，它被设计为捕捉整个输入序列的信息，常用于分类任务，然后将 [CLS] 标记的表示传递给一个全连接层，最终得到分类结果。

## (2) 诈骗摘要生成器

团队为得到诈骗文本分析摘要，向用户展示我们做出判断的原因，增加可信度，并进行针对性改善，我们研发一个专门面向诈骗的大语言模型，判断诈骗的同时，生成风险内容摘要，工作过程如下图5.1：

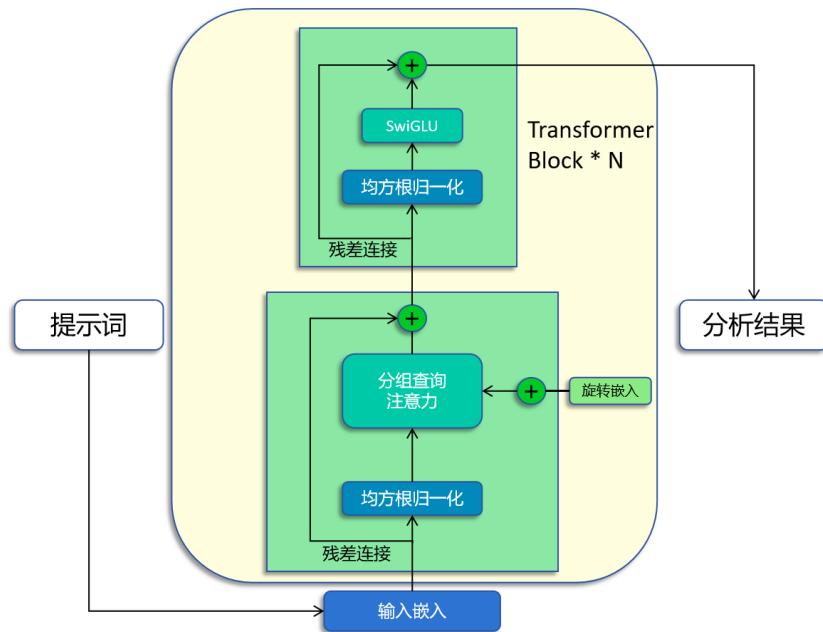


图 3.24 摘要生成器框架图

其中，技术细节如下：

### (a) Decoder-only 架构 [63]

模型的架构是决定其性能、效率和适用性的关键因素，直接影响模型的表达能力、计算效率、泛化能力等多个方面。为了生成诈骗文本摘要分析，我们采用更好的生成文本效果的 Decoder-only 架构的模型。Decoder-only 架构也被称为生成式架构，仅包含解码器部分。Decoder-only 是一种自回归模型，将解码器自己当前步的输出加入下一步的输入，解码器融合所有已经输入的向量来输出下一个向量，所以越往后的输出要考虑更多的输入，很适合完成诈骗信息摘要分析一类的任务。

### (b) 旋转位置编码（RoPE）[64]

位置编码在 Transformer 中具有关键作用，为模型提供了序列中各元素的位置信息，决定在自然语言处理（NLP）时的顺序敏感性，保证关键信息的语句通顺、逻辑性强。我们在位置编码时引入一种复数的思想，通过旋转向量来引入位置信息。我们删除了传统的绝对位置嵌入，并在网络的每一层增加了 RoPE。RoPE 通过绝对位置编码的方式实现了相对位置编码，综合了绝对位置编码和相对位置编码的优点。

---

此外使用旋转位置嵌入保持了序列长度的灵活性和随相对距离的增加而衰减的 token 间依赖性。而为了进一步加强模型检测长的诈骗文本能力，我们在 self-attention 自注意力层中的 q,k,v 向量都加入了位置信息，如下式：

$$\begin{aligned} q_m &= f_q(x_m, m) \\ k_n &= f_k(x_n, n) \\ v_n &= f_v(x_n, n) \end{aligned} \tag{3.25}$$

$q_m$  表示第  $m$  个 token 对应的词向量  $x_m$  集成位置信息  $m$  之后的 query 查询向量，而  $k_n$ 、 $v_n$  则分别表示第  $n$  个 token 对应的词向量  $x_n$  集成位置信息  $n$  之后的 key 键向量、value 值向量。

位置编码过程中，不同位置的元素会被赋予不同的编码，使得模型在计算注意力时可以考虑到位置差异。位置信息与输入特征（ $q$ 、 $k$ 、 $v$  向量）的结合，使每个位置的特征表示更丰富、更具辨识度，使模型在捕捉序列中的模式时更具优势，特别是在需要注意到局部或全局模式时。

### (c) 分组查询注意力 (Group Query Attention, GQA[65])

查询注意力机制在深度学习模型中具有重要意义，能够有效捕捉长距离依赖关系，提高计算效率，提升模型表现。而多查询注意力机制 MQA(Multi-Query Attention)[66] 虽然能最大程度减少 KV Cache 所需的缓存空间，但是其精度同时也大打折扣。因此，我们采用 GQA 机制，其中 Q 依然是多头，但是分组共享 K,V，既减少了 K,V 缓存所需的缓存空间，也暴露了大部分参数不至于精度损失严重。GQA 机制如下：

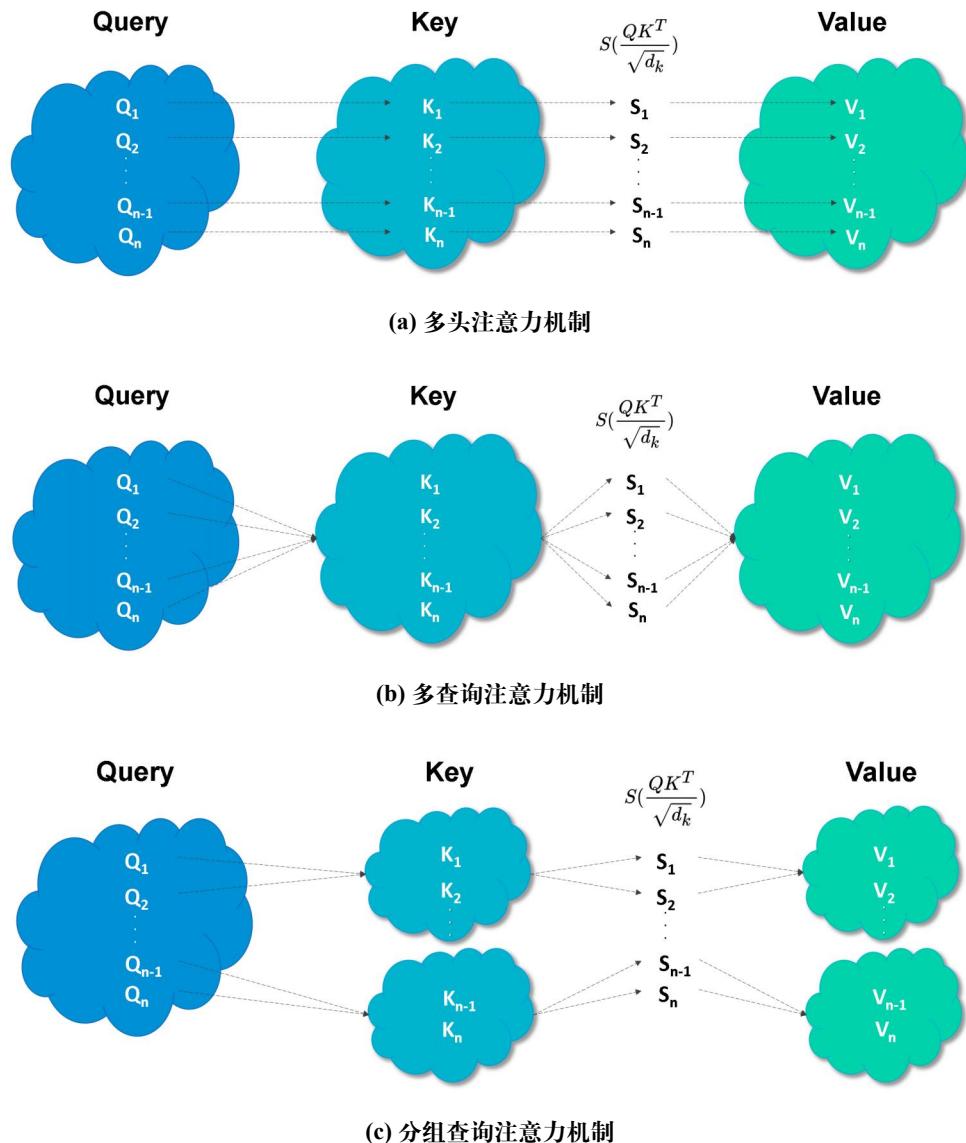


图 3.25 注意力机制

#### (d) 均方根层归一化 (RMSNorm[67])

归一化技术在模型训练当中起着至关重要的作用，可以加速模型的收敛，提高训练速度。目前许多的 Transformer 模型归一化的方法都是层归一化 (LayerNorm[68])。然而，LayerNorm 每次需要计算均值和方差，增大了运算开销。因此，我们采用 RMSNorm 技术，省去了计算样本与均值的差的过程，相当于仅使用  $x$  的均方根来对输入进行归一化，因而具备更快的训练速度。公式如下：

$$\bar{x}_i = \frac{x_i}{RMS(x)} gi \quad RMS(x) = \sqrt{\frac{1}{H} \sum_{i=1}^H x_i^2} \quad (3.26)$$

### (3) 生成器微调

利用 LoRa (Low-Rank Adaptation of LLMs) [69], 我们对全参数微调的增量参数矩阵  $\Delta W$  进行低秩分解近似表示, 即对参数做降维处理, 以此来只对一个参数量更小的矩阵进行低秩近似训练, 将  $\Delta W$  表示为两个参数量更小的矩阵 A 和 B, 如下式:

$$W_0 + \Delta W = W_0 + BA \quad (3.27)$$

其中  $B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times d}$  为 LoRA 低秩适应的权重矩阵, 秩  $r$  远小于  $d$ , 此时, 微调的参数量从原来  $\Delta W$  的  $d \times d$ , 变成了 B 和 A 的  $2 \times r \times d$  这在很大程度上减少了参数训练量, 我们通过此方法可以快速地对大模型进行训练微调。如图3.26:

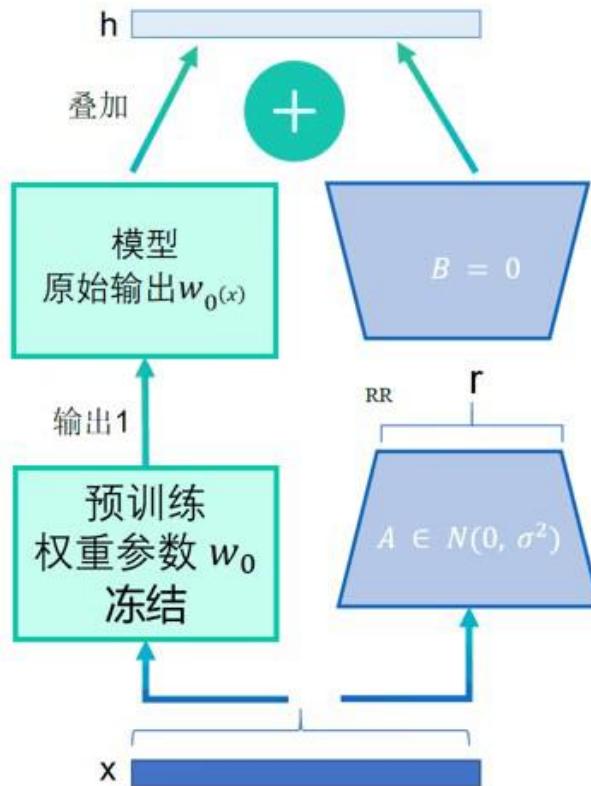


图 3.26 模型低秩分解

模型训练后, 我们进行模型的参数合并, 以此来完整地呈现模型的能力。我们把更新后的矩阵 A,B 与原参数进行原始参数 W 进行合并, 具体过程如图3.27:

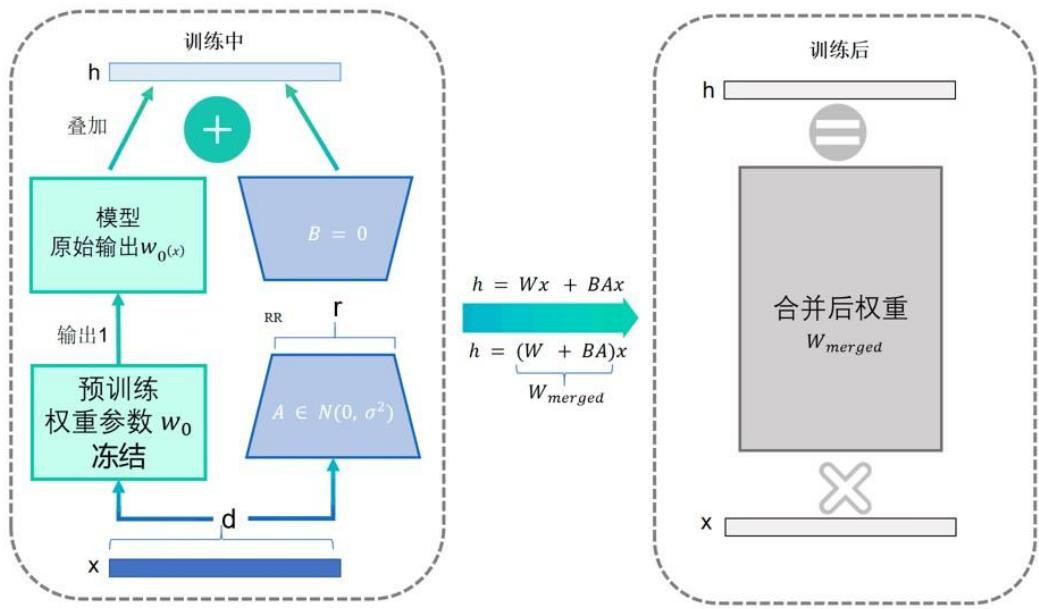


图 3.27 模型参数合并

#### (4) 生成器量化

我们将模型的量化划分为以下步骤：

- ① 使用校准数据运行模型前向传播，收集每层的输入输出。
- ② 对每层的权重应用量化函数，计算量化误差。
- ③ 使用拉格朗日乘子法和海森矩阵的逆来调整权重，最小化量化误差。

其中，我们利用量化函数将权重映射到离散值集合，同时最小化量化误差。

我们使用校准数据用于确定量化参数，如缩放因子和零点偏移。通过分析模型在这些数据上的行为，可以更好地理解权重分布和激活模式，从而选择最合适的量化策略。而后，我们利用校准数据的输入输出构建每层的海森矩阵（Hessian matrix），即损失函数关于权重的二阶偏导数矩阵，用于调整权重以最小化量化误差。特别的，在量化过程中，不断人为或自动地调整权重以补偿量化引入的误差。校准数据提供了上下文信息，帮助量化算法确定如何调整权重以最小化整体的量化误差。量化算法如下：

---

**Algorithm 2** 量化算法: 在已知海森逆矩阵  $H^{-1} = (2XX^T + \lambda I)^{-1}$  和权重矩阵  $W$  的块大小  $B$  的条件下量化权重矩阵  $W$

---

```
1:  $Q \leftarrow 0_{d_{row} \times d_{col}}$                                 // 量化输出
2:  $E \leftarrow 0_{d_{row} \times B}$                                 // 矩阵块量化误差
3:  $H^{-1} \leftarrow \text{Cholesky}(H^{-1})^T$                 // 海森逆矩阵
4: for  $i = 0, B, 2B, \dots$  do
5:   for  $j = i, \dots, i + B - 1$  do
6:      $Q_{:,j} \leftarrow \text{quant}(W_{:,j})$                       // 量化列
7:      $E_{:,j-i} \leftarrow (W_{:,j} - Q_{:,j})/[H^{-1}]_{jj}$       // 计算量化误差
8:      $W_{:,j:(i+B)} \leftarrow W_{:,j:(i+B)} - E_{:,j-i} \cdot H_{j,j:(i+B)}^{-1}$  // 在块中更新权重
9:   end for
10:   $W_{:,i:(i+B)} \leftarrow W_{:,i:(i+B)} - E \cdot H_{i:(i+B),i:(i+B)}^{-1}$  // 更新所有剩余权重
11: end for
```

---

最后，我们再利用校准数据来验证量化模型的性能。通过在这些数据上评估量化模型，可以确保量化后的模型在关键任务上仍然保持足够的精度。

### 3.2.4 基于大数据技术的用户风险形象分析

当前，电信网络诈骗已突破传统广撒网模式，逐步演变为依托深度伪造、社会工程学与大数据分析的精准化攻击。诈骗分子通过多维度数据挖掘构建受害者数字画像，利用伪造身份信息、定制化话术及实时交互策略制造沉浸式诈骗场景，使得单一维度的实时拦截技术难以应对持续演变的长期风险。用户历史行为中潜藏的脆弱性特征往往分布于异构数据源中，传统风控系统因缺乏跨场景数据融合能力与深度关联分析手段，无法有效识别此类隐蔽风险信号。

为构建主动式风险防控体系，“一盾当关”通过深度挖掘用户通信行为特征与历史风险轨迹，打造动态演进的用户风险画像。该系统融合通话情感识别、行为序列建模与大数据关联分析技术，不仅解析通话时长、异常中断等高危行为信号，更运用自然语言处理精准捕捉“转账”“验证码”等风险关键词，结合知识图谱实现诈骗案例的智能匹配与预警。通过 TF-IDF 算法量化文本风险值、反应延迟模型量化决策迟疑度，形成覆盖“数据采集-特征提取-风险评估-知识推送”的全链路防护机制，在 Apache Spark 实时计算框架与 GDPR 合规体系的支撑下，为每位用户构筑起兼具精准识别力与隐私安全性的智能防御屏障。

---

首先，通过分析用户的通话数据，系统能够提取出多维度的信息来构建用户的风险画像。这些数据包括但不限于通话时长、通话内容、通话情感，以及用户在通话过程中的行为特征。系统自动提取通话内容、通话行为检测结果，汇总成易骗特征分析。此外，如果用户在特定时间段内突然中断通话或急速挂断电话，系统将把这些行为标记为风险信号；当这些行为与通话音频情感分析结果（如恐慌或焦虑情绪，频繁的语速加快、声音颤抖等情感特征）相匹配时，风险等级将进一步提高。此外，用户的反应模式也会被纳入分析范围，比如迟疑、犹豫不决等情绪状态，往往预示着用户在面对潜在诈骗时的决策困难，这些行为特征能够帮助系统识别出需要更多关注的高风险情境。这些反应时间可以通过反应延迟时间公式来计算：

$$\Delta t = t_{\text{end}} - t_{\text{start}} \quad (3.28)$$

其中， $\Delta t$  为反应延迟时间， $t_{\text{end}}$  为用户做出决定或反应的时间， $t_{\text{start}}$  为用户开始表现迟疑的时间。较长的反应延迟时间通常表明用户在面对潜在诈骗时的决策迟疑，是识别诈骗时的一种防范信号。系统将其记录在用户的个性化风险形象中，为用户防诈提供帮助。

其次，基于风险摘要报告数据，系统能够进一步分析用户过往的诈骗经历，从而更加精准地描绘用户的风险形象。通过对历史诈骗数据的整合，系统能够识别出用户常遭遇的诈骗类型以及几乎未接触过的诈骗类别。例如，如果某用户频繁接到涉及假冒银行工作人员的诈骗电话，系统将标记此类诈骗为该用户的高风险领域。相反，如果某个用户几乎没有遭遇过投资诈骗，那么该用户在未来遇到此类诈骗时的反应模式可能与其过往经验有所不同。历史数据的汇总为用户风险画像的完善提供了重要依据，系统可以根据这些信息预测用户在未来遇到类似诈骗时的应对态度和风险暴露程度。

大数据技术在实现用户风险形象分析中，主要通过两大功能模块来支撑：特征提取与数据搜索。首先，在特征提取方面，系统利用大数据技术对用户的通话数据与历史诈骗数据进行分析，提取出与诈骗行为相关的特征词和行为模式。TF-IDF 用来提取语音转文本后的关键词，检测文本中的异常词汇和潜在的诈骗术语。对于词频：

$$TF(t, d) = \frac{\text{词项 } t \text{ 在文档 } d \text{ 中出现的次数}}{\text{文档 } d \text{ 中所有词项的总数}} \quad (3.29)$$

对于逆文档频率：

$$IDF(t) = \log \frac{N}{df(t)} \quad (3.30)$$

---

其中， $N$  是语料库中文档的总数， $df(t)$  是包含词项  $t$  的文档数。TF-IDF 分数是上述两者的乘积，可以帮助识别出在诈骗电话中常见的诈骗关键词。例如，若通话文本中频繁出现“急需转账”、“身份证件信息”这些高风险词汇，系统会根据这些特征词过滤掉与用户实际风险匹配的诈骗类别。通过这种方式，系统能够自动识别与用户风险画像匹配的诈骗类型，进而为用户展示精准的特征提示。

其次，基于大数据技术的数据搜索功能能够帮助系统根据用户的特征画像进行更为精准的诈骗案例检索和知识推送。通过数据挖掘和知识图谱技术，系统可以为用户提供与其风险画像相关的诈骗案例和防范知识。知识图谱构建依赖于图的连接结构，通过图算法来执行数据搜索和匹配。当用户的风险形象发生变化时，系统会自动在海量的诈骗案例库中进行检索，寻找相似的诈骗案例并推送给用户。与此同时，基于用户的风险特征，系统还可以根据大数据分析提供个性化的安全建议，如提示用户如何在诈骗电话中识别伪装、如何防范新的诈骗手法等。

为安全、合适地存放和处理用户的隐私数据，一盾当关使用 Apache Hadoop 存储和处理历史数据，包括用户的通话记录、诈骗案例等。使用 Apache Spark 进行实时数据流处理，分析用户当前的通话数据并即时识别潜在的诈骗风险。在合规性方面，系统严格遵守 GDPR、CCPA 等隐私保护法律法规，实施数据匿名化和最小化数据收集的原则，确保用户的个人信息不被滥用。

## 第 4 章 系统实现

### 4.1 前端架构

前端架构不仅决定了前端部分的代码结构和开发方式，还影响着系统的可维护性、扩展性和性能。在具体实现方面，团队选择了 Vue3 和 Django 作为前后端开发的技术栈，通过实现前后端独立开发和部署，使得两个框架能够充分发挥各自的优势，从而为 Web 应用程序的开发提供了强大的支持。同时，Nginx 作为反向代理和静态文件服务器，为平台的功能实现提供了高性能和良好的可扩展性。

Vue3 的核心运行机制如图4.1所示，图中展示了其组件渲染和响应式更新的过程。当组件的状态发生变化时，Vue3 的响应式系统会触发重新渲染。具体过程如下：首先，组件的数据通过 getter 函数进行访问，Vue3 通过该访问行为收集数据依赖项，确定需要观察的数据。数据被收集后，Vue3 会创建一个 Watcher 观察者来监控数据变化。当数据变化时，setter 函数被触发，Vue3 会标记出依赖项并通知观察者。接着，观察者接收到通知后，触发组件的重新渲染。在渲染过程中，Vue3 根据组件的状态生成新的虚拟

DOM 树，随后进行高效的 DOM 更新操作。整个过程保证了 Vue3 应用的界面与数据状态始终保持同步，并且优化了性能。

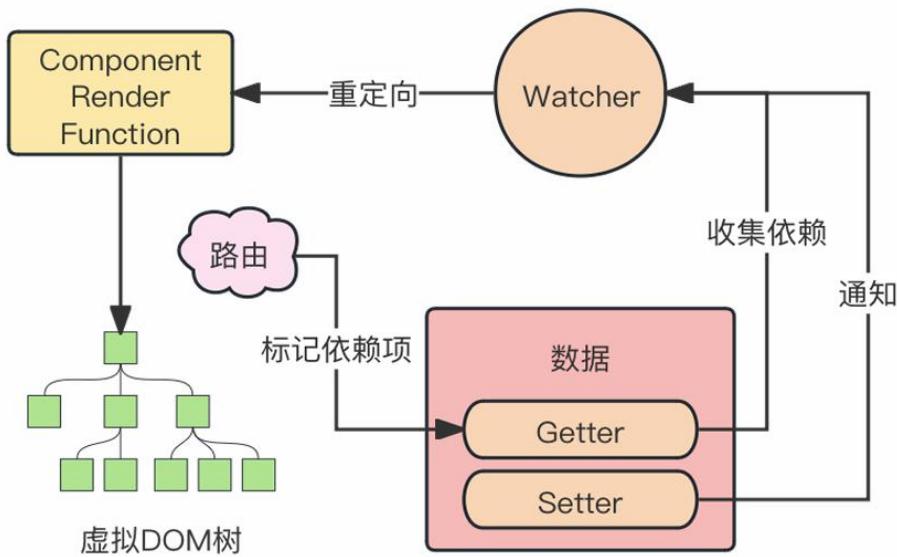


图 4.1 Vue3 核心运行机制

Web 服务器响应用户的请求流程如图4.2所示：

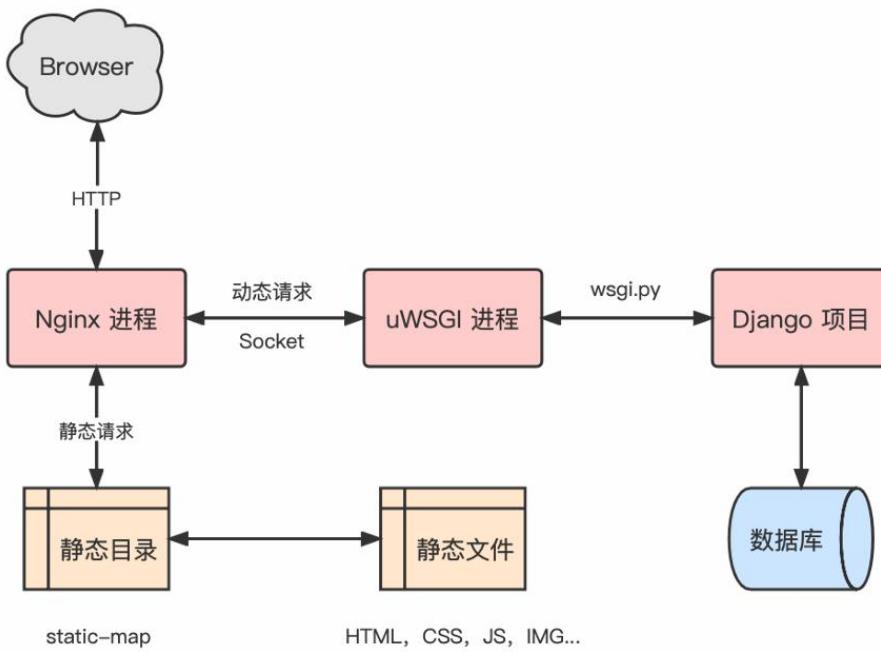


图 4.2 PC 端系统响应客户请求流程图

用户通过客户端（如浏览器）登录系统，并向服务器发出请求（Request）以获取资源。Nginx 作为系统的外部代理服务接口，负责接收客户端发送的 HTTP 请求，并对请求进行解包和分析。对于静态资源请求，系统根据 Nginx 配置文件中定义的静态文件目

录映射规则，返回相应的资源。在配置文件中，可以设置多个映射规则，将特定的 URL 路径指向服务器上的不同静态文件目录。

在本系统的部署配置中，项目的静态文件存放在远程服务器上，并通过 Nginx 代理来提供访问。这些静态资源的请求通过 Nginx 的配置文件指向远程服务器上的静态文件目录，Nginx 会返回相应的资源。对于动态内容的请求，Nginx 使用 Socket 通信将请求转发给 uWSGI 服务器。uWSGI 进一步将请求传递给 WSGI。WSGI 是 Web 服务器与应用程序之间的标准接口。WSGI 接口根据请求调用 Django 框架中的模型层文件和函数来处理相应的请求。模型处理完请求后，生成修复后的图像，并将结果回传给 WSGI。WSGI 负责将处理后的数据打包后发送回 Web 服务器。最终，uWSGI 接收到处理后的数据，并将其转发给 Nginx 代理服务器，Nginx 则将响应结果返回给客户端。

Django 中各部件处理请求的过程如图4.3所示。当用户从浏览器发出请求时，中间件（middlewares）负责对请求进行预处理。接着，URLconf 根据映射找到相应的视图函数，视图函数通过模型（models）访问底层数据。本系统通过 manager 实现与数据库的交互，而模板（templates）则负责页面的渲染，最终将渲染结果返回给视图函数。

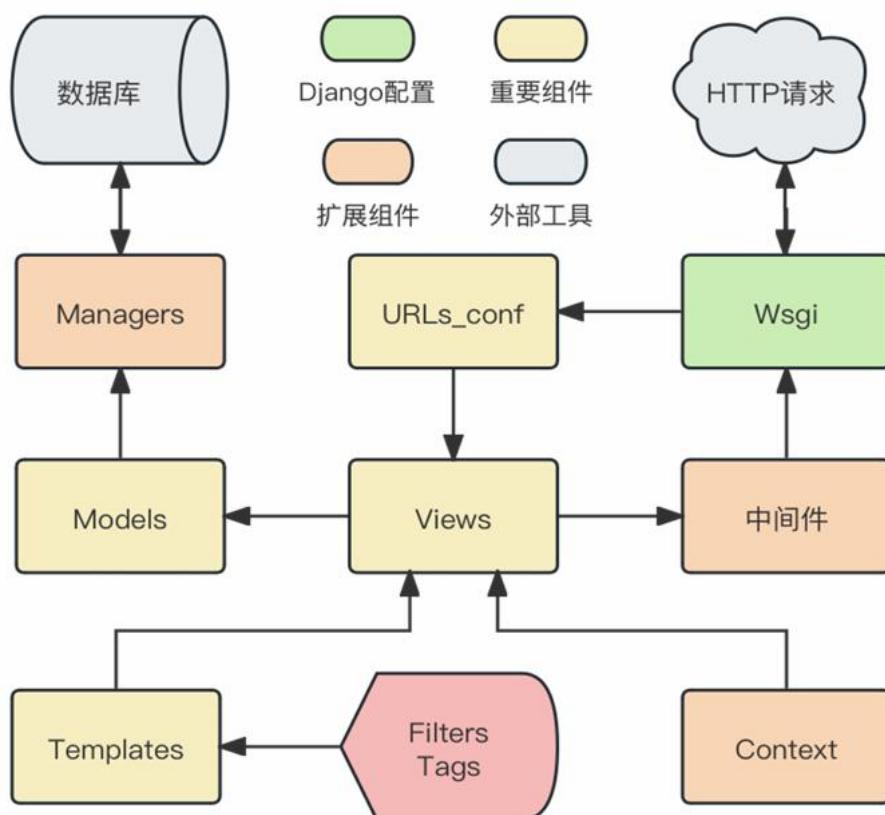


图 4.3 后端 Django 处理框架

- URLs\_conf: 负责 URL 配置，Django 使用它来解析请求的 URL，并确定应该由哪个

视图处理该请求。

- Views: 视图层负责处理业务逻辑，并与模型交互以获取或存储数据。
- Models: 模型层处理与数据库的所有交互，定义了数据结构，并提供了与数据库交互的方法。
- Managers: 作为模型的一部分，封装了数据库查询操作，简化了数据访问。
- Database: 用于存储伪造的视频数据。
- Middlewares: 在请求/响应过程中执行会话管理、用户认证等功能。
- Templates: 模板层用于生成 HTML 响应，支持使用 Filters 和 Tags 处理数据展示逻辑。
- Context: 上下文用于将额外的数据传递给模板，以便在生成 HTML 时使用。

## 4.2 系统部署

系统的应用部署对于确保其功能和效益的实现至关重要，在涉及安全和数据保护的领域更是如此。合理的部署不仅为系统的全面功能提供坚实基础，还能有效保障用户的隐私安全，防止未经授权的访问和数据泄露等攻击行为。同时，优秀的部署方式提高了系统的可靠性和稳定性，使其在处理大量数据和并发检测请求时表现出色。通过注重安全、隐私和易用性，我们的系统能够为用户提供一个安全可靠且易于使用的解决方案。

系统部署图如图4.4:

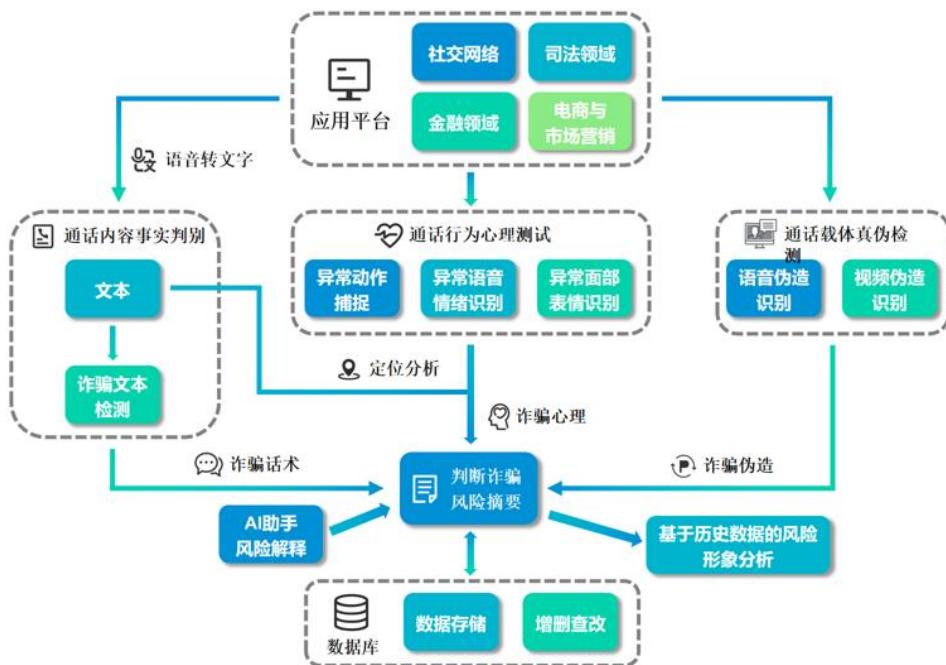


图 4.4 系统部署实现图

由于多模态技术的发展，诈骗细节会广泛存在于通话内容、通话载体、通话行为中。

---

为确保可以做到细致且全面地检测通话过程中的诈骗痕迹，系统将从通话内容、载体、行为三个方面进行分析，并判断诈骗的发生，以及提供风险摘要内容。通过风险形象与AI助手对用户的个性化风险进行动态评估与干预。

①通话内容的检测部分会以音频转换后的文本作为输入，使用基于诈骗检测任务优化后的大语言模型对文本进行分析处理，输出诈骗判断，并将识别发现的疑似诈骗话术返回到风险摘要中。

②通话载体的检测部分会以视频、音频内容作为输入，利用视频伪造检测、音频伪造检测技术对视音频内容进行处理，检测其中是否存在由伪造技术重新合成的内容，输出诈骗判断，并返回伪造内容到风险摘要中。

③通话行为的检测部分会以视频、音频内容作为输入，利用动作捕捉、面部表情分析、语音情绪分析对音视频进行跨模态联合处理，得到各个异常情绪时间点。系统将会重点关注多个模态下异常情绪交叉的时间点，并结合对应时间段的文本内容，分析得出诈骗心理。最后输出诈骗判断并返回识别到的异常动作和异常面部、语音情绪到风险摘要内容中。

④风险形象与AI助手部分通过综合分析用户的行为、情感和环境等多个维度的数据，不仅能够及时识别潜在的诈骗风险，还能通过智能预警和干预机制，帮助用户有效规避诈骗行为。这一模块使得系统能够根据用户的具体情况，提供更为精确和个性化的防范措施，进一步提升了系统对复杂风险的应对能力。

系统组件之间进行数据传递和综合分析，完成各自功能后，将所得结果进行相互交换，以此来完成上述功能。系统包含组件及其功能见表4.1：

---

表 4.1 系统各组件及功能

类别	系统组件	功能实现
内容	音频转换器	利用语音识别技术，将输入的语音信号转换为可编辑和处理的文本内容。
	诈骗文本检测器	通过训练大语言模型，学习诈骗话术的特征，对输入的文本进行模式匹配和风险评估。
载体	声码器检测器	分析音频信号的频谱特性，检测音频中是否存在非自然生成或修改的痕迹。
	时序不一致性伪造检测器	利用深度学习算法，结合唇动与语音的时序对应关系，识别 Deepfake、LipSync 伪造视频中的不一致性。
行为	语音情绪识别器	采用情感计算技术，分析音频中的语调、语速等特征，判断说话人的情绪状态。
	静态面部表情识别器	基于图像处理技术，识别图像中人物面部表情特征，推断其情绪状态。
	动态面部表情识别器	结合视频帧间信息，跟踪人物面部表情变化，连续评估情绪状态。
风险	异常动作捕捉器	应用动作捕捉技术，分析视频中人物的动作轨迹和姿态，识别不符合常规或预设的异常动作。
	风险形象构建与分析模块	基于用户的多维度数据（如行为、情感、设备等），动态构建用户的风险画像，为安全策略的优化提供支持。
	AI 预警和干预系统	利用 AI 模型对用户行为进行监测，识别风险状态并及时触发警告，提供相应的干预措施，如信息提示、行为引导等。

## 4.3 模块训练

### 4.3.1 数据来源

(1) 通话载体模块经过我们团队深入调研，目前互联网上公开的深度伪造数据集中，大多数只包含单一的视频或图像素材，尚未有专门针对 LipSync 伪造检测而设计的音视频数据集。为了解决音频数据不足的问题，我们联合实验室原创了一个高质量的音视频伪造数据集，如4.5。该数据集汇集了通过几种顶尖 LipSync 算法生成的 12,000 个视频，并且具备灵活扩展能力，理论上可动态扩展至 60 万个具有时序特征的样本。我们对数据进行音频降噪的数据预处理，场景覆盖率公共数据集和真实的微信视频通话。

为了更好地模拟真实环境的细微差别，我们采用了六种不同程度的扰动技术来构建数据集，即饱和度、对比度、压缩、高斯噪声、高斯模糊和像素化，从而模拟现实场景中可能会产生的质量损失，保证数据集的真实性。



图 4.5 AVLips 数据集构建

(2) 通话行为模块异常动作的识别是本系统诈骗检测模块中的关键环节。然而，通过调研我们发现，当前学术界和工业界在此方向尚缺乏针对性强、适用于诈骗行为识别的公开数据集，现有的动作识别数据往往难以准确覆盖诈骗过程中可能出现的异常行为特征。为填补这一空白，我们构建了一个专门面向诈骗场景的异常动作图像数据集，涵盖约 8,000 张高质量图像，部分样例如图 4.6 所示。

为了使数据集更贴近真实应用场景，我们引入了多种视觉扰动因素进行样本增强，包括“彩色光源干扰”、“低照度环境”以及“复杂背景混淆”等，以模拟各类诈骗行为可能发生的真实场所与环境。此外，为提升模型的鲁棒性与泛化能力，我们还引入了部分现有的动作识别数据集，与自建数据集进行联合训练，从而有效增强了系统在多样化环境中的识别能力。

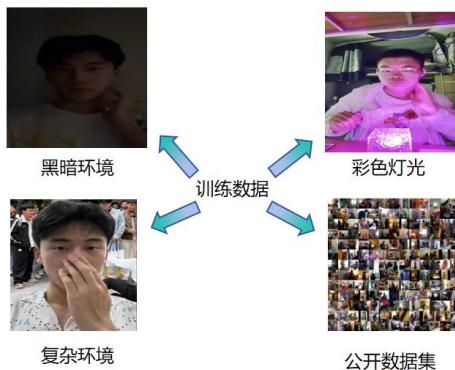


图 4.6 诈骗异常动作数据集构建

(3) 通话内容模块诈骗文本的识别是本系统风险感知模块的核心功能之一。然而，通过调研我们发现，目前学术界和工业界尚缺乏专门针对诈骗语料的高质量文本数据

集，尤其在文本类型细分与风险分层标注方面存在显著不足。为弥补这一空白，我们自主构建了一个面向多类诈骗场景的高质量文本数据集，如??。其中，共收录 8,497 条诈骗文本样本，覆盖“中奖通知”、“虚假贷款”、“冒充客服”、“投资诈骗”、“冒充公检法”等多种真实高发诈骗类型，具备较强的代表性与实际应用价值。

在数据构建过程中，我们不仅为每条文本标注了详细的诈骗类型，还引入了“高风险 (high)”、“中风险 (medium)”与“低风险 (low)”的风险分层标签体系，便于系统在识别后按风险等级进行处置。同时，为提升系统在实际环境下的鲁棒性与泛化能力，我们对原始文本样本进行了多种语言增强处理，包括同音字替换、随机字符删除、常用词错写扰动等，从而模拟真实诈骗信息中常见的规避检测手法。最终形成的数据集不仅具备丰富的诈骗表达样式，也更贴近现实中复杂多变的语言特征，为模型训练与评估提供了坚实支持。



图 4.7 诈骗话术文本数据集构建

#### 4.3.2 数据训练

针对本系统的三个核心模块——通话载体模块、通话行为模块与通话内容模块，我们设计了系统化且差异化的数据训练流程，确保每一模块均可在其特定任务中实现最佳性能。

在通话载体模块中，我们使用自建的 AVLips 伪造数据集作为训练核心，重点聚焦于音视频同步伪造检测任务。训练前，首先对音频部分进行统一的降噪处理，视频部分则引入饱和度、对比度、压缩、高斯噪声、高斯模糊与像素化等多种扰动策略，以增强模型在不同画质下的鲁棒性。模型训练采用多尺度时序特征提取机制，融合了音频与视频双模态输入，利用跨模态一致性作为伪造判别的主要依据。训练过程中，采用对比损失与分类损失联合优化策略，有效提升模型对细微伪造特征的感知能力。

在通话行为模块中，针对异常动作识别任务，我们使用自建的诈骗异常动作图像数据集，并结合公开动作识别数据集进行联合训练。在预处理阶段，我们引入了彩色光源扰动、低照度、复杂背景等视觉干扰因素，增强样本多样性。训练过程中，模型采用图

---

像级卷积特征提取结合时序建模结构，能够识别静态异常姿态与动态异常行为。同时，为提升模型对少样本类别的识别能力，我们引入了迁移学习与少样本学习框架，通过冻结预训练模型部分层权重，再利用诈骗动作数据进行微调，使模型更快适应目标场景。

在通话内容模块中，我们构建了多类诈骗话术文本数据集，并配套设计了三层风险分级标签体系。训练过程中，模型采用文本分类框架，基于预训练语言模型（如 BERT）进行微调，并结合文本增强方法（如同音替换、错字扰动等）提升模型对规避检测话术的识别能力。为强化模型在多类别与多层次风险判断中的表现，我们引入了多任务学习机制，使模型在进行诈骗类型分类的同时，也能输出风险等级预测结果，从而提高系统整体识别的准确性与精度。

综上，本系统在数据训练阶段，针对不同模块的任务目标，采取了针对性的建模与优化策略，通过多模态、多扰动、多层次标签设计，实现了对真实诈骗场景的高拟合与强泛化能力，为系统在实际应用中稳定高效运行提供了有力支撑。

#### 4.3.3 改进过程

在系统实际部署阶段，我们在原有训练流程的基础上，进一步进行了多轮迭代优化，特别围绕数据集的可拓展性、场景贴合度与伪装鲁棒性三个核心维度展开了改进工作。与现有同类数据相比，我们在数据构建、特征设计与技术集成方面实现了显著的技术增量，进一步增强了系统在真实诈骗通话场景下的适应能力和准确率。

在通话载体模块方面，针对已有深度伪造检测数据集（如 DFDC、FaceForensics++）大多集中于视频维度、忽视音视频时序一致性的局限，我们对 AVLips 数据集进行了二次扩展。新增的伪造样本涵盖了更多元的话术内容与表情变换类型，特别引入了微信视频通话真实语料作为伪造参考源，使合成样本更具自然感和干扰性。在技术实现上，我们引入了基于 Transformer 架构的跨模态一致性编码器（Cross-modal Alignment Transformer, CAT），用于深度建模语音与唇动之间的微时差，从而显著提升了对微伪造视频的检测能力。

在通话行为模块方面，我们改进了原始数据集在动态行为表达不足的短板。在部署过程中，结合行为检测模型与视频关键帧提取策略，我们从真实诈骗通话中抽取了更多含异常动作的动态图像序列，并扩充至帧间时序样本。同时，我们在原有扰动增强策略基础上，增加了“设备抖动模拟”、“摄像头遮挡模拟”等新型扰动模式，使模型在面对真实用户低清通话、光线干扰等复杂环境下依旧保持较高识别能力。此外，我们引入时序卷积网络（TCN）替代传统静态 CNN 结构，以更好地建模连续动作趋势，解决突发异常行为难以捕捉的问题。

---

在通话内容模块方面，我们在原有诈骗文本数据集的基础上，通过联动真实举报数据平台与安全实验室，增补了多个当前正在高发的诈骗类型（如 AI 换脸诈骗、虚拟客服引导式诈骗）。这一增量极大拓展了模型对新型诈骗手法的识别能力。在风险标注体系上，我们也从原始三层风险等级扩展为五层细化评分体系，引入“行为意图评分”与“语言引导强度”两个新指标，使模型不仅能判断文本是否为诈骗，还能评估其诱导性强度与诈骗成熟度。在技术方面，我们升级为多通道注意力机制的文本编码器，并引入知识增强模块（incorporated knowledge embedding）结合已有诈骗术语库进行上下文语义补全，从而有效识别规避型、隐喻型和间接性诈骗话术。

总体来看，相较于已有公开数据与模型，我们系统的数据构建与训练过程在真实度、扩展性和攻击对抗性三个方面都进行了深度增强与技术叠代，特别是在模拟真实通话场景中的细节干扰与跨模态一致性建模方面具有明显领先优势。这些技术增量使系统在实际部署过程中，面对真实诈骗攻击时展现出更强的鲁棒性与响应能力，也为后续的智能化反诈系统发展提供了坚实基础。

## 4.4 系统展示

本系统的主要功能页面分为实时通话检测和离线检测两部分，旨在提供全面的安全防护。实时通话检测功能的设计初衷是在保证用户进行正常通话的同时，实时识别和防范潜在的诈骗风险，确保用户的安全。该功能广泛适用于商务洽谈、亲友联络和在线交易等场景，帮助用户实时识别和阻止诈骗行为，从而避免经济损失和个人信息泄露。离线检测功能则满足了用户在非实时通话时的安全需求，允许用户对已录制或存储的通话内容和文件数据进行分析，以识别潜在的诈骗风险。这一功能适用于回顾重要通话、调查取证和文件检测等场景，确保用户在任何时候都能对通话内容进行全面的安全检查。通过结合实时检测和离线检测功能，系统构建了一个多层次的安全防护体系，不仅提高了整体安全性，还显著提升了用户体验，使用户在使用过程中能够更加安心、便捷地防范诈骗行为。

### 4.4.1 实时通话检测

实时通话检测页面是我们系统的核心功能之一，旨在通过先进的技术手段，确保音频通话和视频通话的安全性和有效性。本页面提供实时监控和分析工具，帮助用户识别并应对潜在的欺诈行为、恶意活动或其他违规行为。同时，我们还提供了录音录屏按钮，用户可得到对应文件进一步分析。

我们先进行了微信实时视频通话测试，如下图所示4.8。

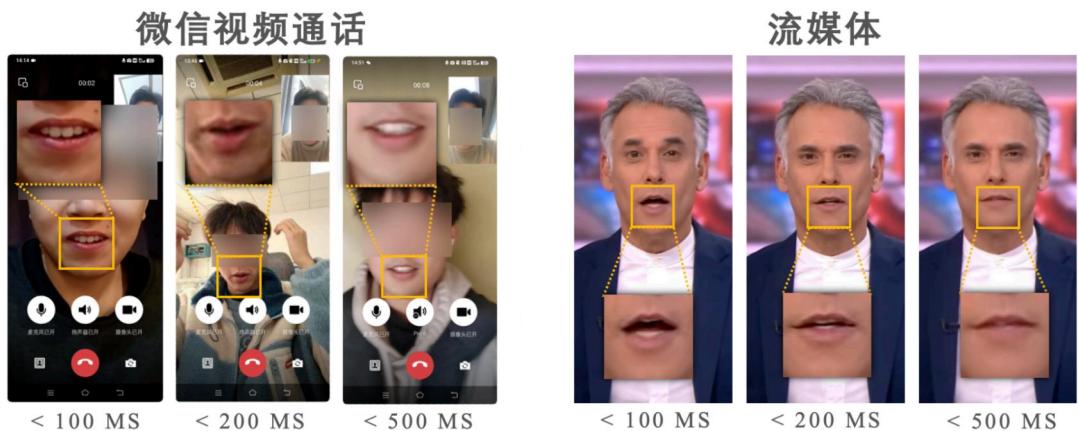


图 4.8 微信通话测试

用户通过输入与对方约定好的房间号进入对应的音视频通话房间，即可正常使用音视频通话的基本功能。同时，我们的系统是支持多人音视频通话的，用户上下滑动即可看到其他用户。



图 4.9 音视频通话

在音视频通话场景下，诈骗分子可能通过 ManyCam、VCam、OBS 等虚拟摄像头软件调用事先伪造好的视频或实时伪造的视频流，伪造合成音频并结合各种话术实施针对性的诈骗。

这里以 ManyCam 为例，诈骗分子上传提前录好的图片或视频到虚拟摄像头，在进行音视频通话的时候选择虚拟摄像头来充当自己的实际摄像头，从而实施诈骗。如图4.10可以看到，诈骗这可以提前准备多段视频，可以做到无缝转换。当然，这也仅仅是诈骗分子实施诈骗的一种方式。

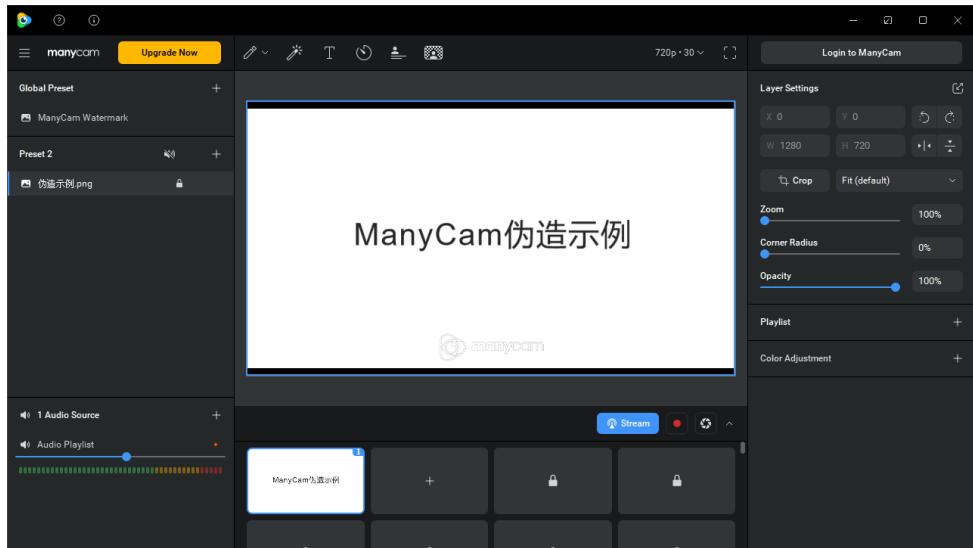


图 4.10 虚拟摄像头软件实施诈骗示例

因此，为了更好的提醒和保护用户，系统提供了实时检测功能。用户点击“实时检测”按钮后，系统会捕捉当前网页的画面和声音发送给后端进行分析，接收返回结果并显示。用户可以点击“查看分析结果”，获取实时的检测结果，并提供下载和复制功能。检测结果如图4.12。

另外，用户点击“开始录屏”与“开始录音”按钮，可分别使用录音录屏的功能，该功能可以与实时检测功能同时使用，用户点击“停止录音”或挂断，即可得到对应的mp4文件和wav文件，文件可用于离线文件检测进一步分析检测。

#### 4.4.2 离线文件检测

用户可以自由选择上传视频、音频、图片、文本文件进行检测，分析后会返回对应的结果与风险摘要。



图 4.11 风险摘要示例（以音频文件检测为例）

风险摘要主要包括是否为诈骗、概要、具体分析三个部分，具体分析又分为音视频是否伪造、文本分析结果、语音转换结果、语音情绪分析、异常动作捕捉和时序图六个部分，详细具体，方便用户阅读。

#### 4.4.3 风险智能分析

##### (1) 摘要报告

系统完成多模态分析后，最终生成的摘要报告是用户了解通话过程中风险信息的核心内容。该报告汇集了系统在通话内容、载体、行为等多个维度的分析结果，并以简洁明了的形式呈现给用户，帮助用户快速了解潜在风险并做出相应决策。生成的摘要报告包括：

- 诈骗判断结果：总结整个通话中是否存在诈骗行为的结论，基于多模态分析的综合结果，提供明确的风险等级和诈骗可能性评估。
- 疑似诈骗话术：从通话内容中提取出与诈骗相关的语句或话术，并标明其出现的具体时间段，便于用户精准识别。
- 伪造内容提示：对通话载体中的视频或音频部分进行检测后，若发现伪造或篡改的痕迹，该部分会列出可疑的伪造内容，并提供详细说明，帮助用户辨识。

- 行为异常分析：分析通话过程中可能存在的异常行为，包括异常的情绪波动、面部表情变化和动作轨迹。通过结合多模态数据，系统能够识别出潜在的诈骗心理活动。

在报告的末尾，系统将根据整个通话的风险评估给出风险等级，并针对当前状况提供后续防范建议。对于诈骗可能性较高的情况，报告会建议立即采取行动，如挂断通话、进行身份验证或向相关部门报告。

生成的诈骗风险摘要报告内容大致如图4.12左侧，从通话载体、通话行为、通话语义三个角度进行诈骗概率的分析，并给出诈骗语义分析的结果，最后根据诈骗载体和诈骗对象的行为进行分析并给出示例以警示用户。



图 4.12 诈骗风险摘要内容

## (2) AI 助手

在系统中，AI 助手作为核心智能分析模块之一，扮演着用户与系统之间的桥梁角色。其接入市面上流行的 DeepSeek-R1 API，基于强大的大数据分析、机器学习以及情感计算技术，能够帮助系统在通话过程中智能分析多模态数据，并生成个性化安全建议。通过与系统的多模态分析接口对接，AI 助手不仅负责为用户提供实时、智能化的辅助，还能根据系统的多模态分析结果生成个性化安全建议，提升用户的安全防范意识，并协助用户快速应对潜在的风险。该模块结合大数据分析、机器学习和情感计算等技术，能在多种复杂情境下对诈骗行为做出精准的预测与响应。

并且，AI 助手能够通过不断的用户交互与数据反馈进行自我学习和优化。随着时

间的推移，AI 助手能够更精准地了解用户的行为特征、情感变化以及安全偏好，进一步提升智能分析和风险判断的准确性。这使得 AI 助手不仅是一个被动的风险提醒工具，更是一个智能化的个人安全助手，能够在复杂多变的风险环境中提供持续、精准的防护。部署效果如图4.12右侧。

### (3) 状态时间图

状态时间图是系统中的一种可视化展示方式，位于首页的右上角，用于直观展示通话过程中潜在诈骗嫌疑的变化趋势。图中的每个点代表在特定时间段内对通话内容、载体和行为的综合分析结果，数值越高，说明该时间段内诈骗嫌疑的概率越大。通过这种方式，用户可以清晰地识别出通话过程中的高风险时间段，快速定位可能的诈骗行为。

状态时间图通过一个曲线图呈现通话期间的风险变化，横轴表示通话的时间，纵轴表示诈骗嫌疑的概率值。图中的峰值部分标明了诈骗嫌疑最大的时间点，而低谷部分则代表嫌疑较小的时段。用户可以通过对图表的观察，快速识别出在某些时段是否存在异常或高风险。



图 4.13 状态时间图

为了提升用户的交互体验，状态时间图支持鼠标悬浮功能。当鼠标悬停在图表的某个具体时间点时，系统将展示该时间段内的详细检测提示。这些提示将结合诈骗检测的各项指标，如通话内容中的可疑话术、异常的情感波动、伪造的音视频信号等，提供具体的分析报告，帮助用户了解该时间段内诈骗嫌疑的来源。状态时间图不仅帮助用户识别通话中的高风险时段，还能结合其他系统模块进行综合分析。例如，在高风险时段，系统可能会提示用户进行身份验证、挂断电话或采取其他防范措施。通过状态时间图的

反馈，AI 助手还可以为用户提供定制化的安全建议，帮助其做出更为及时和理智的决策。

#### 4.4.4 用户风险形象分析

系统将基于用户的历史通话数据以及诈骗风险摘要报告，结合大数据分析技术和智能算法，深入挖掘和识别潜在的电信诈骗风险。通过对用户过往通话内容的语义分析，系统能够精准提取涉及诈骗的关键特征，如特定的诈骗关键词、可疑语境、异常通话模式等，从而归纳出诈骗分子惯用的手法、策略以及潜在的作案趋势。结合这些分析结果，系统将对用户的受骗风险进行多维度画像建模，以量化评估用户在不同诈骗类型下的风险暴露程度，并据此提供针对性的防诈策略建议。



图 4.14 用户风险形象及防诈知识智能推送

为了提升用户的安全防范意识，系统将智能推送与用户高频接触的诈骗类型相关的防诈知识和真实案例解析，使用户能够深入理解诈骗手法的运作模式，掌握有效的应对策略。同时，为了构建更完善的防护体系，系统不仅关注用户已遭遇或高风险的诈骗类型，还会主动推送用户较少接触或尚未遭遇的诈骗类型的相关信息，帮助用户建立全面的安全防范意识，做到未雨绸缪，提升整体防诈能力。

此外，系统将结合用户的行为特征，动态调整推送内容，确保防诈信息既精准匹配

用户需求，又具有前瞻性和广泛性，从而实现精准化、个性化的反诈骗教育。通过构建多层次、全方位的防诈安全体系，系统能够在提升用户认知能力的同时，提高其在面对复杂诈骗手段时的警觉性和应对能力，从而有效降低受骗风险，构建更加安全、可信的通信环境。

#### 4.4.5 信息交流论坛

为向用户提供更加丰富、真实且实用的诈骗案例及防诈知识，本系统特设“诈骗社区动态”论坛功能。该论坛旨在为用户搭建一个交流互动的平台，使其能够分享自身的受骗经历、讨论各类诈骗手法，并总结学习防诈知识的经验与心得，从而提高个人防范意识，构建更具安全性的网络环境。

系统依托先进的关键词过滤与语义特征提取技术，能够智能分析用户发言内容，并自动归类至相应诈骗类型，同时为发言打上精准的类别标签。这一功能不仅便于其他用户直观地识别诈骗案例及防诈知识，还能有效提升信息检索的效率。用户可通过点击发言下方的类别标签，快速跳转至同类案例讨论页面，或使用搜索功能精准查找相关内容，系统将自动推送同标签下的优质用户分享与讨论内容，帮助用户全面了解诈骗手法，提高警惕，增强防范能力。

The screenshot displays the 'Community Dynamics' section of the system. On the left, there's a vertical sidebar with icons for messaging, community, and other features. The main area has two main sections:

- Community Dynamics:** This section shows three user posts:
  - 云淡风清:** A post about a landlord scam where the user was tricked into giving a deposit.
  - (^ω^)暖青:** A post about a scammer who pretended to be in an accident and asked for money.
  - 海阔天空:** A post encouraging users to share their stories.
- Anti-fraud Knowledge Q&A:** This section contains a question and several options for answers:

以下哪种行为最容易被网络诈骗分子利用？

  - A. 在社交平台上公开分享个人旅行计划
  - B. 在购物网站上使用信用卡支付
  - C. 在公共Wi-Fi环境下登录银行账户
  - D. 在收到陌生短信后点击链接并填写个人信息
- Recent High-frequency Fraud Cases:** This section includes a case analysis and a feature prompt:

**案例分析**

**刷单诈骗** 系统近期监测到多起刷单诈骗案件，犯罪分子通常会通过社交平台、短信或电话，声称“动动手指就能轻松赚钱”，诱导受害者下载指定的App或加入刷单群。受害者完成首次刷单后可能会收到小额返利，犯罪分子利用这种“正反馈”逐渐建立信任，随后要求受害者进行大额刷单或支付“解冻费”、“保证金”等，最终在受害者支付高额款项后直接拉黑或失联。

**特征提示:**

  - 出现“高额回报”、“轻松赚钱”等宣传语时务必提高警惕。
  - 正规客服从不通过私人电话或非官方渠道联系用户。
  - 官方客服不会要求下载任何第三方定位应用。

**安全建议:**

遇到类似情况，建议立刻挂断电话，切勿透露个人信息。如已涉及财产损失，请第一时间联系银行冻结账户并报警。下载国家反诈中心App，实时接收反诈预警和最新诈骗动态。

图 4.15 用户信息交流动态论坛

---

## 第5章 测试分析

### 5.1 系统性能测试

为了测试系统相关性能，我们从通话内容、通话载体、通话行为三个方面针对每个板块分别进行有效性测试和鲁棒性测试。有效性测试的目的在于测试本项目设计的方法是否能够应用于诈骗防御问题，效果是否良好；鲁棒性测试目的在于检测我们的模型是否具有较强的抗干扰能力，是否能在更贴近真实的使用场景中取得良好表现。最后，我们对各个模型设置评判标准，并总结分析测试结果。

#### 5.1.1 通话内容

##### · 有效性测试

为检测通话内容中的诈骗文本检测器能否很好地识别诈骗分子的诈骗话术，我们设计了这次测试。

测试时，我们使用自制的诈骗话术数据集，数据集包含正常内容、诈骗内容等文字数据。

我们自制数据集的原因在于考虑到市面上现有的数据集不够复杂，往往以较短的垃圾信息为主，故本团队独立制作了一个包含 8497 条信息的数据集。数据集涵盖了 14 种信息，分别属于‘正常内容’，‘中奖’，‘网络刷单’，‘虚假招聘’，‘网络交易’，‘绑架’，‘代购’，‘网络交友’，‘信用贷款’，‘炒股理财’，‘冒充政府人员’，‘保险’，‘学术圈诈骗’。数据集信息如下表5.1：

表 5.1 诈骗数据类型及数量

类型	数量	训练集	测试集
正常内容	2053	1442	611
中奖	202	145	57
网络刷单	832	578	254
虚假招聘	523	369	154
网络交易	689	476	213
绑架	214	151	63
代购	199	143	56
网络交友	223	148	75

接下页

表 5.1 – 续

类型	数量	训练集	测试集
信用贷款	560	404	156
炒股理财	563	389	174
冒充政府人员	365	257	108
保险	194	137	57
学术圈诈骗	1880	1325	555
合计	8497	5964	2533

我们将数据集随机打乱，并按约 7: 3 的比例划分为训练集和验证集，在自制的诈骗信息数据集上使用两种模型进行分类，将模型预测的类别和实际类别进行比较以评估我们的模型能否有效解决通话内容方面的诈骗问题。测试流程如下图所示5.1：

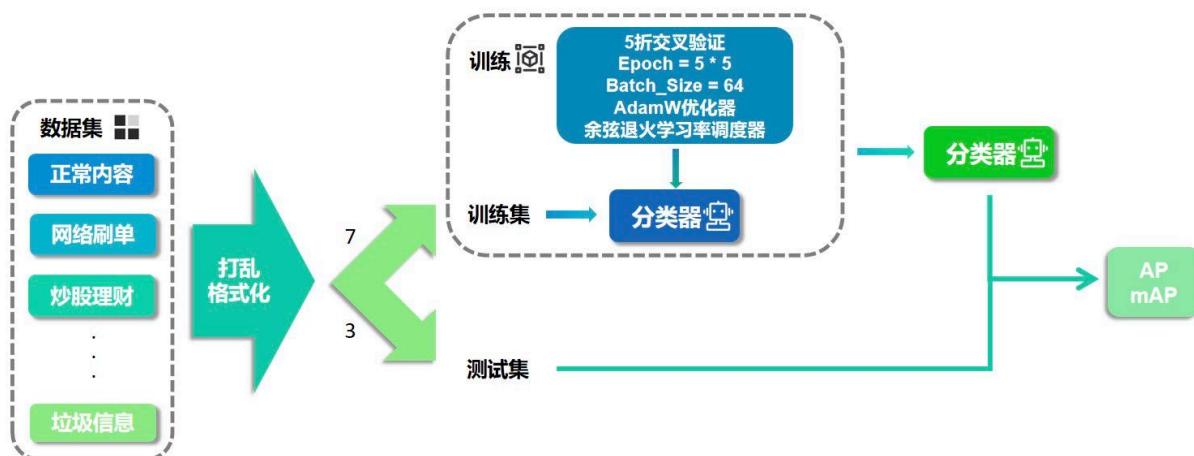


图 5.1 通话内容检测的有效性测试方案

我们使用 AP 和 mAP 两项指标来评估测试结果。AP 是指在不同的召回率下，计算精度的平均值。具体来说，AP 是精度-召回曲线（PR 曲线）下的面积，可以通过计算每个类别的精度-召回曲线下的面积来得到。PR 曲线越靠近右上角，AP 值越高，表示模型在该类别上的表现越好；mAP 是指所有类别的 AP 的平均值。它综合考虑了所有类别的检测性能，是一个整体的性能评价指标，可以通过先计算每个类别的 AP，然后对所有类别的 AP 取平均值得到。

通过这两个指标，我们不仅能够评估模型在某些类别上的表现是否良好，而是评估在所有类别上的表现是否良好，可以直观地反应我们模型的效果。

---

测试结果如下表5.2

表 5.2 通话内容检测的有效性测试结果

指标	诈骗类型						
	正常内容	中奖	网络刷单	假冒领导、熟人	虚假招聘	买卖虚拟货币	绑架
AP(%)	99.71	98.28	78.95	92.60	99.17	73.34	93.83

我们的模型在各种诈骗类型上的表现非常出色，展示了其高效的检测能力和强大的鲁棒性。在所有诈骗类型中，模型的 AP 均保持在较高水平，尤其是在“正常内容”、“中奖”、“虚假招聘”、“保险”和“垃圾信息”等类别中，精度接近或超过 98%。

总体而言，模型的 mAP 为 90.35%，这表明模型在所有类别上的综合表现非常优秀。无论是面对复杂的诈骗信息还是不同类型的扰动攻击，模型都能保持稳定的性能。这些结果充分证明了我们模型的准确性和可靠性，为其在实际应用中的有效性提供了有力支持。

### · 鲁棒性测试

我们设计这次测试以检测我们的模型在面对一系列干扰下是否还能保持良好性能。

我们对上文使用的同样的测试集集进行了 4 种方式的扰乱攻击，分别是：在 0.2 的概率下对每个字进行随机同音字替换、在 0.4 的概率下对每个字进行随机同音字替换、在 0.3 的概率下进行随机字删除、在 0.3 的概率下随机置换邻近字。

表 5.3 扰动攻击

攻击类型	攻击概率
随机同音字替换	0.2、0.4
随机字删除	0.3
随机置换邻近字	0.3

其中，**随机置换邻近字、随机字删除**：随机置换邻近字可能不会导致整个句子结构的破坏，而随机字删除虽然会减少信息量，但句子的基本结构仍然可以保持，这可能使得模型能够更容易地恢复或推断出原始意图；**同音字替换**：同音字替换虽然保留了发

音，但可能会改变句子的语义，导致模型难以正确理解句子的真实意图。然后我们在添加扰乱的数据集上进行测试。我们仍然使用 AP 和 mAP 两项指标来评估测试结果。

最终，检测效果如下表5.4：

**表 5.4 通话内容检测的鲁棒性检测结果**

指标	诈骗类型	扰动攻击方式			
		同音字替换（概率 0.2）	同音字替换（概率 0.4）	随机置换邻近字	随机字删除
AP(%)	正常内容	98.84	98.09	99.38	98.19
	中奖	98.28	88.53	98.28	98.28
	网络刷单	77.58	67.73	76.05	78.53
	假冒领导、熟人	86.89	80.13	92.00	91.43
	虚假招聘	95.94	86.51	96.76	95.14
	买卖虚拟货币	69.17	60.34	67.55	71.97
	绑架	92.36	85.16	88.03	90.73
	代购	94.81	86.38	98.25	98.25
	网络交友	83.59	77.33	85.04	80.85
	信用贷款	75.70	64.00	76.30	78.09
	炒股理财	80.31	71.44	84.40	84.97
	冒充政府人员	88.45	81.62	91.10	89.31
	保险	96.56	89.75	96.61	98.28
	垃圾信息	99.03	94.59	99.40	98.67
mAP(%)		88.39	80.83	89.22	89.48

我们的模型在各种扰动攻击方式下表现出色，展示了其强大的鲁棒性和高效的检测能力。无论是同音字替换、随机置换邻近字还是随机字删除，模型在大多数诈骗类型上的 AP 都保持在较高水平，尤其是在“中奖”、“保险”“冒充政府人员”、“假冒领导、熟人”等类别中，精度几乎没有显著下降。

总体而言，模型在无扰动情况下的 mAP 为 90.35%，在不同扰动攻击方式下的 mAP 也都保持在 80% 以上，最高达到 89.48%。这些结果表明，我们的模型不仅能够准确识别各种诈骗信息，还能在面对不同类型的扰动攻击时保持稳定的性能。这些测试结果充分证明了我们模型的鲁棒性和可靠性，为其在实际应用中的有效性提供了有力支持。

### 5.1.2 通话行为

检测对象的通话行为包括静态和动态面部表情识别器和异常动作捕捉器，

#### (1) 面部表情识别器

依据通话行为的不同，我们的面部表情技术测试分为动态面部识别表情测试和静态

---

面部表情识别测试。

### (a) 动态面部表情识别测试

#### · 有效性测试

我们在 FERV39k 和 MAFW 数据集上对模型进行测试，并与其他模型进行对比，综合评估结果。

FERV39k 数据集包含 38935 个野外视频片段，是目前最大的野外 DFER 数据集。所有的视频片段收集自 4 个场景，这些场景可以进一步划分为 22 个细粒度的场景，如犯罪、日常生活、演讲和战争。每个片段由 30 个单独的注释者注释，并分配给七个基本表达之一作为 DFEW。我们将所有场景的视频片段随机打乱，分成训练 (80%) 和测试 (20%) 两个部分，它们间没有重叠。

MAFW 数据集包含 10045 个真实的视频片段，是第一个具有多种情感类别的大规模情感数据库。具体来说，该数据集包含 11 个单表情类别，包括愤怒、厌恶、恐惧、快乐、悲伤、惊讶、轻蔑、焦虑、无助、失望和中立。此外，它还包含 32 个多表情类别，捕捉了更复杂的情感状态和情绪组合。除了这些视频片段，MAFW 数据集还提供了丰富的情感描述文本，为情感分析和识别提供了多模态、多标签的研究基础。我们采用 5-fold 交叉检验进行模型的测试。

我们选择 DPCNet[70]、CLIPER[46]、M3DFEL[71]、ResNet18-ViT、AEN[72] 等模型作为我们的基线模型。其中，DPCNet 以创新的双路径结构、自适应融合策略和轻量化的设计，在多个数据集上表现优异；CLIPER 模型以零样本学习的能力尤其突出，在动态面部情绪识别方面准确率超过了许多传统的模型；M3DFEL 模型，即“Multi-modal 3D Feature Learning”，拥有三维特征学习的时间信息处理能力；AEN 模型，即“Attention-Enhanced Network”，通过引入注意力机制使其在许多数据集上表现优异。

最终，测试流程图如下 5.2：

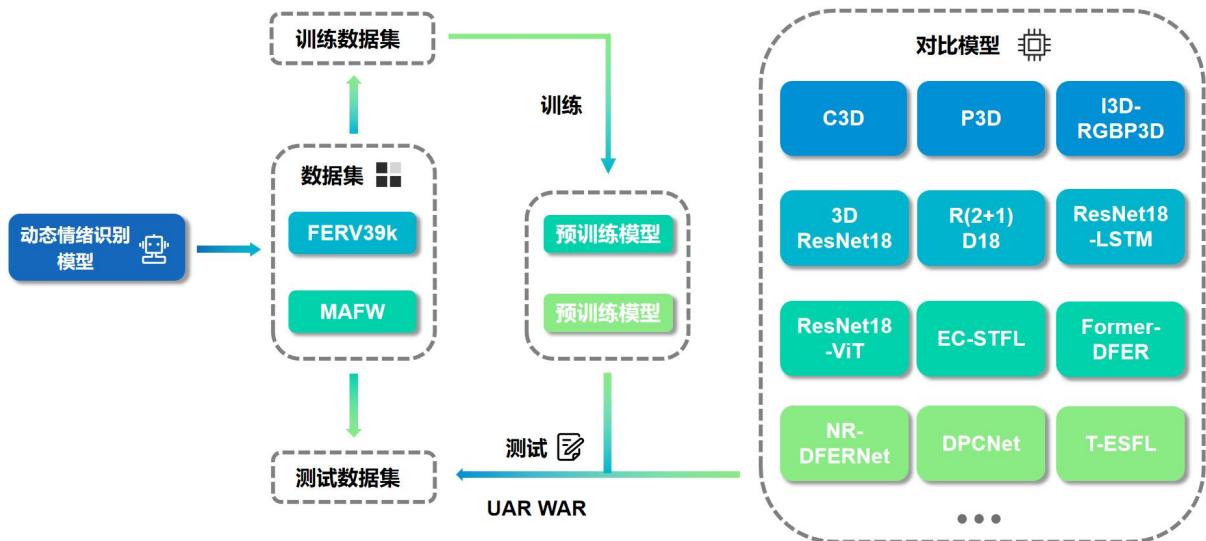


图 5.2 动态表情的有效性测试方案

我们使用 UAR 和 WAR 两项指标来评估测试结果。

UAR 即 Unweighted Average Recall，未加权平均召回率。在评估模型在各个类别上的整体性能比较常用，特别是当数据集中的类别分布不均匀时。它是在不考虑每个类别的样本数量情况下，通过计算所有类别的召回率的平均值得到的，具体计算方法如下：

$$UAR = \frac{1}{C} \sum_{c=1}^C Recall_c \quad (5.1)$$

WAR 即 Weighted Average Recall，加权平均召回率。他会考虑到每个类别的样本数量，因此可以更好地反映分类器在处理具有不同样本分布的类别时的性能。计算公式如下：

$$WAR = \frac{\sum_{c=1}^C (TP_c \times Recall_c)}{\sum_{c=1}^C TP_c} \quad (5.2)$$

最终，测试结果如下 5.5：

表 5.5 动态表情识别的有效性测试结果

模型	FERV39k		MAFW	
	UAR	WAR	UAR	WAR
C3D	22.68	31.69	31.17	42.25
P3D	23.20	33.39	-	-
I3D-RGB	30.17	38.78	-	-
3D ResNet18	26.67	37.57	-	-
R(2+1)D18	31.55	41.28	-	-
ResNet18-LSTM	30.92	42.95	28.08	39.38
ResNet18-ViT	38.35	48.43	<u>35.80</u>	47.72
EC-STFL	-	-	-	-
Former-DFER	37.20	46.85	31.16	43.27
NR-DFERNNet	33.99	45.97	-	-
DPCNet	-	-	-	-
T-ESFL	-	-	33.28	<u>48.18</u>
EST	-	-	-	-
LOGO-Former	38.22	48.13	-	-
IAL	35.82	48.54	-	-
CLIPER	<u>41.23</u>	<u>51.34</u>	-	-
M3DFEL	35.94	47.67	-	-
AEN	38.18	47.88	-	-
我们的模型	<b>41.27</b>	<b>51.65</b>	<b>39.89</b>	<b>52.55</b>

看表可知，我们的模型在两个数据集上的表现均优于其他模型，展示了其在动态面部表情识别任务中的强大性能和鲁棒性。特别是在 MAFW 数据集上，UAR 和 WAR 均为最高值，表明我们的模型在处理复杂数据时具有较高的准确性和稳定性。具体来说，在 FERV39k 数据集上，我们的模型的 UAR 为 41.27，WAR 为 51.65；在 MAFW 数据集上，UAR 为 39.89，WAR 为 52.55，均为最高值。相比之下，其他模型在这两个数据集上的表现则相对较弱，例如 C3D 在 FERV39k 数据集上的 UAR 为 22.68，WAR 为 31.69；在 MAFW 数据集上的 UAR 为 31.17，WAR 为 42.25。总体而言，这些结果充分证明了我们的模型在动态面部表情识别中的优越性和可靠性。

### · 鲁棒性测试

我们将在涵盖多种极端光线、面部遮挡等复杂现实情况的 DFEW 数据集上将我们的模型与其它模型进行性能对比，其整体流程与有效性测试下的测试类似5.2。

DFEW 数据集包含 11697 个野外视频片段，所有的样本都被分割成 5 个相同大小的部分，没有重叠。每个视频在专业指导下由 10 名标注者单独标注，并分配到 7 种基本表情(即快乐、悲伤、中立、愤怒、惊讶、厌恶和恐惧)中的一种。这些视频剪辑从全球 1500 多部电影中收集，涵盖了各种具有挑战性的干扰，如极端光照、遮挡和不同的头部姿势。我们采用 5 折交叉验证 (5-fold crossvalidation)[73] 作为模型的评估方式。

表 5.6 动态表情识别的鲁棒性测试结果

模型	UAR	WAR
C3D	42.74	53.54
P3D	43.97	54.47
I3D-RGB	43.40	54.27
3D ResNet18	46.52	58.27
R(2+1)D18	42.79	53.22
ResNet18-LSTM	51.32	63.85
ResNet18-ViT	55.76	67.56
EC-STFL	45.35	56.51
Former-DFER	53.69	65.70
NR-DFERNet	54.21	68.19
DPCNet	57.11	66.32
EST	53.94	65.85
LOGO-Former	54.21	66.98
IAL	55.71	69.24
CLIPER	<u>57.56</u>	<u>70.84</u>
M3DFEL	56.10	69.25
AEN	56.66	69.37
我们的模型	<b>56.61</b>	<b>71.25</b>

测试结果中，我们模型相比于效果第二的 CLIPER 模型，在 UAR 和 WAR 两个指标上分别提高了 2.05% 和 0.41%，说明我们的模型在多种极端光线、面部遮挡等复杂现

实情况下依然能展现极高的性能，这充分证明了我们的模型在鲁棒性测试中的优越性和可靠性。

### (b) 静态面部表情识别技术测试

#### · 有效性测试

我们在 RAF-DB 数据集上测试我们的模型，并将我们的模型与设定的基线模型通过 AP 和 mAP 两个指标进行对比分析。

RAF-DB 数据集是一个大规模的 FER 数据集，它整合了来自不同现实生活场景的图像，如社交媒体视觉内容和电影帧，生动地展示了在自然环境中识别表情的复杂性和多样性。该数据集涵盖了七种基本表情以及 21 种复合表情，本实验只使用其中的七种基本表情。数据集被划分为 12,271 张训练图像和 3,068 张测试图像。

我们选择 SSD[74]、RetinaNet[75]、YOLOv3[76]、CenterNet[77]、EfficientNet[78]、YOLO 和 FER-YOLO-Mamba 几个模型作为基线模型。其中，SSD 模型基于前馈神经网络、采用非极大值抑制的方法生成最终的检测结果，在各种数据集上表现优异；RetinaNet 模型通过引入 Focal Loss 解决了类别不平衡的问题，在 COCO 数据集上实现了当时最佳的性能；CenterNet 模型通过定位目标的中心点，提升性能的同时提高了模型的效率，在多种数据集上取得了优异的效果；EfficientNet 利用一种新的复合缩放方法，在分类准确性方面取得了当时的最佳性能。

最终，测试流程图如下 5.3：

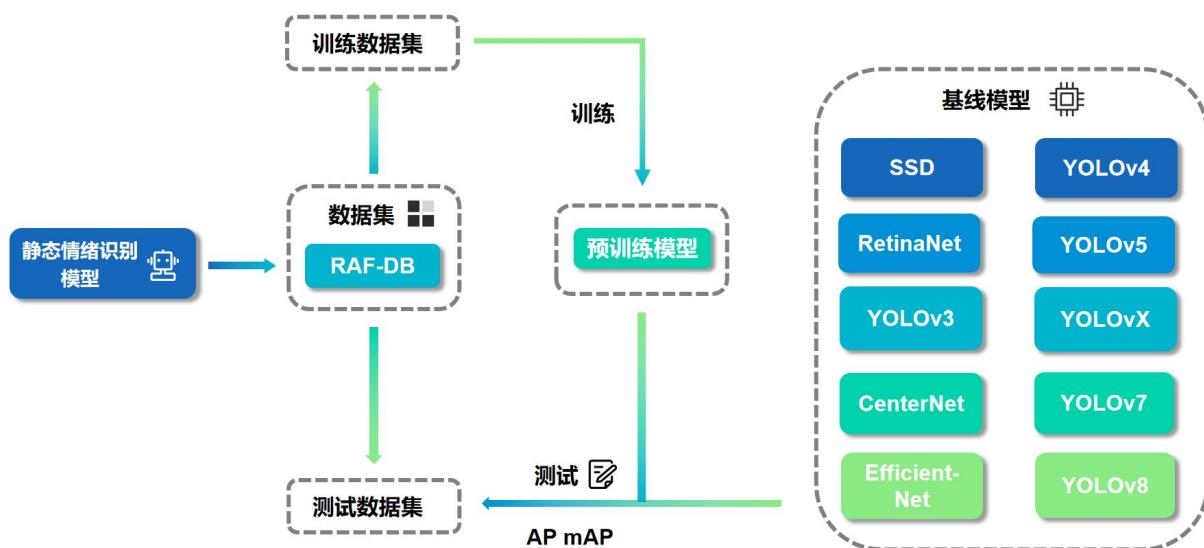


图 5.3 静态表情识别的有效性测试方案

评估指标 AP 与 mAP 的解释见前文 5.1.1。最终，测试结果如下表 5.7：

表 5.7 静态表情识别的有效性测试结果

模型	AP(%)							mAP(%)
	生气	厌恶	害怕	快乐	平静	悲伤	惊讶	
SSD	81.23	62.91	57.01	95.72	80.34	78.32	89.71	77.89
RetinaNet	82.07	53.74	53.56	94.63	80.13	77.50	87.80	75.63
YOLOv3	58.01	28.56	37.02	88.09	67.89	63.78	72.59	59.42
CenterNet	53.26	17.41	30.62	91.32	75.36	66.83	84.63	59.92
EfficientNet	68.72	52.25	45.47	93.75	78.67	76.96	84.31	71.45
YOLOv4	39.25	0.00	10.36	87.61	52.77	45.72	59.91	42.23
YOLOv5	45.75	8.86	0.00	91.77	64.73	65.60	74.36	50.15
YOLOvX	78.38	62.40	57.85	96.82	80.45	83.35	89.56	78.40
YOLOv7	62.20	55.80	44.72	92.01	73.20	74.72	74.57	68.17
YOLOv8	74.50	50.40	50.85	93.33	76.39	76.30	82.89	72.09
我们的模型	79.55	64.32	62.00	97.43	83.23	84.22	91.44	80.31

从表格中可以看出，我们的模型在所有类别上的表现都非常出色，尤其是在“厌恶”、“害怕”、“快乐”、“平静”、“伤心”和“惊讶”类别中，AP 值均为最高或接近最高值。具体来说，我们的模型在“Disgust”类别上的 AP 为 64.32%，在“Fear”类别上的 AP 为 62.00%，在“Happy”类别上的 AP 为 97.43%，在“Neutral”类别上的 AP 为 83.23%，在“Sad”类别上的 AP 为 84.22%，在“Surprise”类别上的 AP 为 91.44%。总体的 mAP 值为 80.31%，也是所有模型中最高的。

相比之下，其他模型的表现则相对较弱。例如，SSD 在“Anger”类别上的 AP 为 81.23%，在“Disgust”类别上的 AP 为 62.91%，在“Happy”类别上的 AP 为 95.72%，在“Surprise”类别上的 AP 为 89.71%，总体的 mAP 值为 77.89%。RetinaNet 在“Anger”类别上的 AP 为 82.07%，在“Happy”类别上的 AP 为 94.63%，总体的 mAP 值为 75.63%。YOLOv3、CenterNet、EfficientNet、YOLOv4、YOLOv5、YOLOvX、YOLOv7 和 YOLOv8 等模型在各类别上的表现也不如我们的模型。

总体而言，这些结果充分证明了我们的模型在动态面部表情识别中的优越性和可靠性。我们的模型在所有类别上的表现均优于其他模型，展示了其强大的检测能力和鲁棒性。

## · 鲁棒性测试

我们在 SFEW 数据集上将我们的模型与基线模型进行对比。SFEW 数据集是专门为研究极端、复杂条件下的 FER 而设计的基准数据集。该数据集的一个显著特点是其“野外”表情的性质，这些表情出现在自然和不受控制的场景中。数据集涵盖 7 种基本表情，样本源自 AFEW 视频数据库，经过精心注释的关键面部表情帧，一共 1,251 张图像，它们描绘了各种光照条件、背景复杂性、头部姿势和面部遮挡等的鲁棒性测试，准确模拟了现实世界表情识别任务中遇到的复杂场景。因此，我们将其选作鲁棒性测试的数据集。

测试流程与有效性测试下的测试类似 5.3。我们仍然使用评估指标 AP 与 mAP，其解释见前文 5.1.1。最终，测试结果如下 5.8：

表 5.8 静态表情识别的鲁棒性测试结果

模型	AP(%)							mAP(%)
	愤怒	厌恶	惊讶	快乐	平静	悲伤	惊讶	
SSD	62.77	47.24	44.74	<b>91.20</b>	<b>55.50</b>	66.48	<u>46.59</u>	59.22
RetinaNet	<u>68.91</u>	58.59	55.23	81.87	43.10	64.16	24.86	56.67
YOLOv3	19.52	0.00	5.88	50.28	37.45	21.11	0.00	19.18
CenterNet	39.54	0.00	25.12	68.57	22.14	43.95	0.00	28.48
EfficientNet	15.40	1.18	17.81	29.68	17.02	29.26	0.72	15.87
YOLOv4	29.58	0.00	0.00	21.12	13.78	21.67	0.00	12.31
YOLOv5	23.56	0.00	0.00	23.64	25.52	22.71	0.00	13.63
YOLOvX	67.01	<b>73.86</b>	<b>70.48</b>	90.81	36.15	<u>70.26</u>	39.55	<u>64.02</u>
YOLOv7	57.47	<u>64.64</u>	52.55	74.34	32.44	48.44	32.26	52.02
YOLOv8	56.24	45.24	53.76	87.50	33.48	44.69	42.68	51.94
我们的模型	<b>74.07</b>	64.49	<u>58.87</u>	<u>90.94</u>	<u>48.01</u>	<b>71.83</b>	<b>58.52</b>	<b>66.67</b>

从表格中可以看出，我们的模型在大多数类别上的表现都非常出色，尤其是在“愤怒”、“悲伤”和“恐惧”类别中，AP 值均为最高或接近最高值。具体来说，我们的模型在“愤怒”类别上的 AP 为 74.07%，在“悲伤”类别上的 AP 为 71.83%，在“惊讶”类别上的 AP 为 58.52%。总体的 mAP 值为 66.67%，在所有模型中排第一。

相比之下，其他模型的表现则相对较弱。例如，SSD 在“快乐”类别上的 AP 为

---

91.20%，在“平静”类别上的 AP 为 55.50%，总体的 mAP 值为 59.22%。RetinaNet 在“愤怒”类别上的 AP 为 68.91%，在“厌恶”类别上的 AP 为 58.59%，总体的 mAP 值为 56.67。其他模型的效果仍然不如我们的模型。

总体而言，这些结果充分证明了我们的模型在动态面部表情识别中的优越性和可靠性。我们的模型在大多数类别上的表现均优于其他模型，展示了其强大的检测能力和鲁棒性。特别是在处理复杂数据时，我们的模型能够保持较高的准确性和稳定性。

### (c) 异常动作

#### · 有效性测试

我们在 Charades 数据集上训练模型。Charades 数据集是一个专注于家庭环境里人类动作捕捉的大规模视频数据集，它收集了人们在执行各种日常活动的视频片段，一共包含 9848 个平均长度为 30 秒的视频样本，涉及在日常对话、手工制作、家庭冲突等 15 种室内场景中的 46 个对象类别的交互，包括大到走路，小到挠头等的 157 个动作类别。数据集的构建考虑到了现实生活中各种场景下的人的行为表现。我们最终提取数据集中视频的关键帧作为模型的输入训练数据，将数据划分为 7986 个训练数据和 1862 个测试数据。

为进一步展现我们模型的性能，我们将模型与其他基准模型进行对比：我们选择 Cascade R-CNN[79]、SDD[80]、DETR[81]、FCOS[82] 几个最先进的模型作为基准模型。其中，Cascade R-CNN 利用多个检测网络按顺序工作，在 CVPR 2018 上发表，成为了当时高精度目标检测领域的里程碑；DETR 即 Detection Transformer，是一种端到端的目标检测模型，它能够直接从图像到边界框和类别标签预测；FCOS，即 Fully Convolutional One-Stage Object Detection，是一种无锚框（anchor-free）的单阶段目标检测模型，能够在许多标准数据集上达到与多阶段检测器相当的水平。

最终，测试流程图如下 5.4：

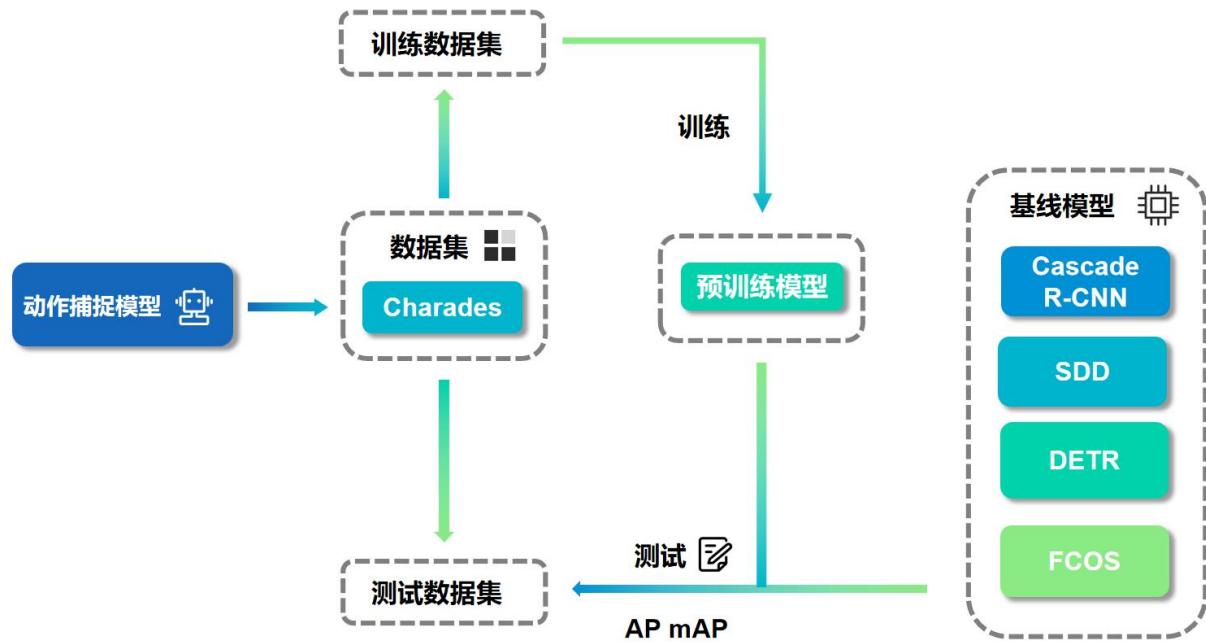


图 5.4 动作捕捉效果图

异常动作捕捉的有效性测试效果图如图5.5：

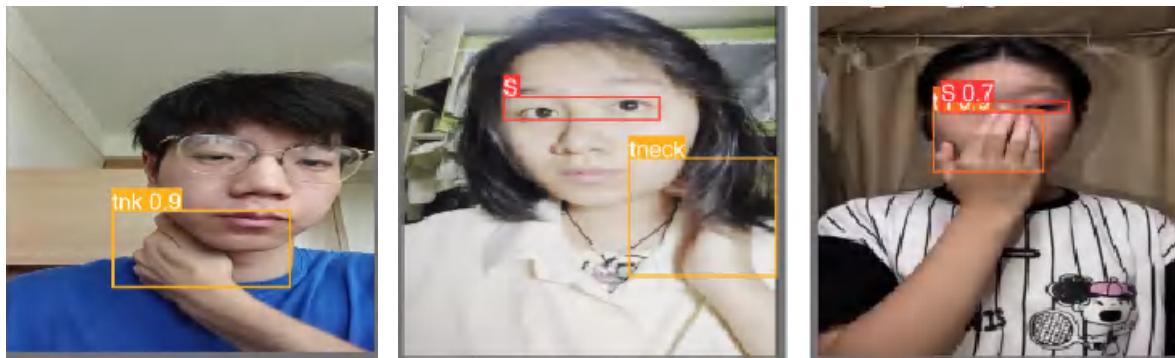


图 5.5 动作捕捉有效性测试效果

我们使用 mAP 和 Mean IoU 两个指标进行对比分析，其中，mAP 指标解释见前文5.1.1；Mean IoU（平均交并比，也称为平均 Jaccard 指数），是一种广泛运用在目标检测和语义分割等领域，评估预测边界框与真实边界框之间的匹配程度的一种指标。Mean IoU 是通过计算所有类别的 IoU（交并比）的平均值得到的。IoU 是评估预测边界框与真实边界框之间匹配程度的一个重要指标。它通过计算预测边界框与真实边界框的交集面积与并集面积的比值，能够准确反映模型的检测精度。IoU 越高，表示预测边界框与真实边界框越接近，模型的检测效果越好。IoU 的计算公式如下：

$$IoU = \frac{|P \cap G|}{|P \cup G|} \quad (5.3)$$

---

其中， $P$  是预测边界框， $G$  是真实边界框。 $P \cap G$  表示预测边界框与真实边界框的交集面积， $P \cup G$  则表示并集面积。则 Mean IoU 的计算公式可以表示为：

$$\text{Mean IoU} = \frac{1}{C} \sum_{c=1}^C \text{IoU}_c \quad (5.4)$$

其中，假设有  $C$  个类别，第  $c$  个类别的 IoU 表示为  $\text{IoU}_c$ 。

通过 Mean IoU，我们能够有效处理类别不平衡的问题。由于它对每个类别的 IoU 进行平均处理，不会因为某些类别样本数量较多而对整体评估结果产生过大影响，从而保证评估结果的鲁棒性。

最终，测试结果如下 5.9：

**表 5.9 异常动作捕捉技术的有效性测试结果**

模型	mAP(%)	Mean IoU(%)
Cascade R-CNN	22.68	31.69
SDD	23.20	33.39
DETR	30.17	38.78
FCOS	26.67	37.57
我们的模型	<b>41.27</b>	<b>51.65</b>

根据测试结果，我们的模型在 mAP 和 Mean IoU 两个指标上均高于其他模型。其中 mAP 达到了 64.78%，相比第二名的 Cascade R-CNN 高出了 1.33%，说明我们模型的动作捕捉精确度和准确性已经超越了目前一些最先进的模型；Mean IoU 指标达到了 67.23%，相比第二名高出了 2.13%，说明我们模型在捕捉动作时的定位精准度也达到了极高的水平。

### · 鲁棒性测试

我们在 HVU 数据集上与基线模型进行对比。HVU (Holistic Video Understanding) 数据集，即全面视频理解数据集，是一个非常大规模的数据集，一共包含 57 万多个视频，9 百万个标注，3142 各类别，包括场景、目标、动作、时间、属性、概念等多个分类任务。由于视频场景多种多样，它能够展现出复杂现实情况下各种光线极端、人体遮挡等的鲁棒性测试。在我们的测试中，只关注动作分类任务，它包含 479568 个视频，245868 个标注和 739 种类别。我们将动作分类任务的部分从数据集中分离出来，同样只关注视频中的关键帧作为模型输入，并将数据集按 9:1 的比例分为训练集和测试集。

鲁棒性的测试流程与有效性测试下的测试类似5.4。我们仍然使用 mAP5.1.1 和 Mean IoU5.1.2 作为指标进行评估。

异常动作捕捉的鲁棒性测试效果图如图5.6：

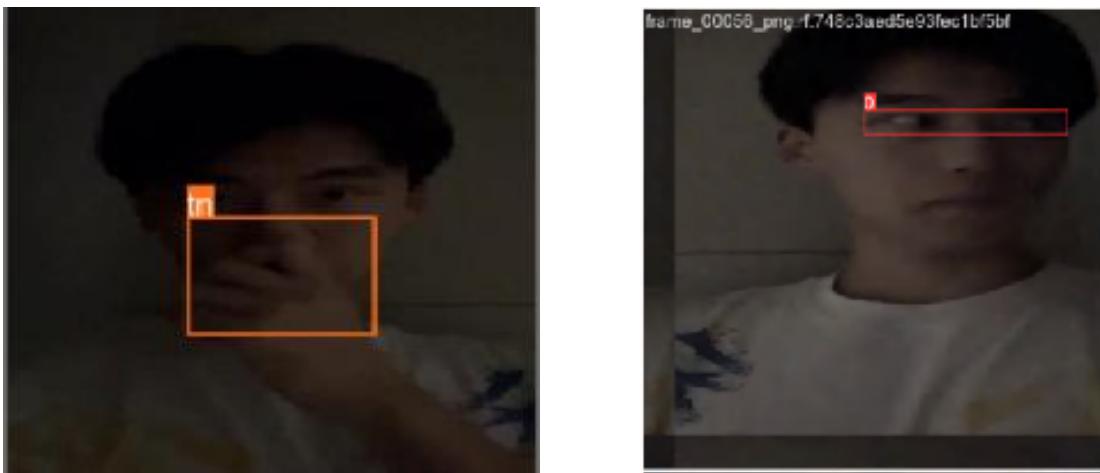


图 5.6 动作捕捉鲁棒性测试效果

最终，测试结果如下表5.10：

表 5.10 异常动作捕捉技术的鲁棒性测试结果

模型	mAP(%)	Mean IoU(%)
Cascade R-CNN	22.68	31.69
SDD	23.20	33.39
DETR	30.17	38.78
FCOS	26.67	37.57
我们的模型	<b>41.27</b>	<b>51.65</b>

根据测试结果，我们在 mAP 和 Mean IoU 指标上依然取得了非常好的结果。mAP 指标上我们仅次于 Cascade R-CNN 模型，Mean IoU 指标比第二名高出了 2.09%。说明我们的模型在复杂的极端现实条件下依然能够达到非常好的识别准确性和动作定位精度。

### 5.1.3 通话载体

#### (1) 时序不一致性伪造检测器

对于时序不一致性伪造检测器测试将在有效性测试和鲁棒性测试下分别进行测试。

##### · 有效性测试

我们挑选了涵盖包括 MakeItTalk[83]、DeepFake、Face2Face[84] 等多种主流的假脸生成算法生成的伪造视频作为测试集，共包含 14500 个样本。

我们将其按 10: 1 的比例分为训练集和测试集，并经过地标检测、唇部裁剪、多帧检测后将训练集送入神经网络训练。然后将训练好的网络与基线模型在数据集上进行准确率、平均精度、假阳率、假阴率等指标的性能对比。

为了更好的突出模型性能，我们挑选了 CVit[85]、DoubleStream[86]、SelfBlended[87]、RealForensics[88] 和 LipForensics[89] 作为基线模型做对比。其中，CVit 模型，即 Convolutional Vision Transformer，结合了卷积神经网络 (CNN) 和 Transformer 的优势，在 ImageNet 数据集上取得了 87.7% 的第一名的准确率；DoubleStream 模型，能够利用图像中不同视图间的信息差异来检测伪造；UniversalDetect 模型，是一种面向开放世界的通用目标检测模型，具备强大的开放世界泛化能力；SelfBlended 模型，通过生成自混合图像来检测伪造，在人工智能顶级会议 CVPR 2022 上获得了口头报告的荣誉；RealForensics 模型，通过自我监督的方式进行训练来检测伪造，在视频压缩或损坏的情况下，会实现非常好的效果；LipForensics 模型，会特别针对人脸视频中的唇部区域进行分析，能够很好的抓住唇部与音频的不一致性来判断视频是否伪造。

通过这些模型进行对比，我们能够很好地判断本项目模型的性能效果。

最终，测试过程如下图5.7：

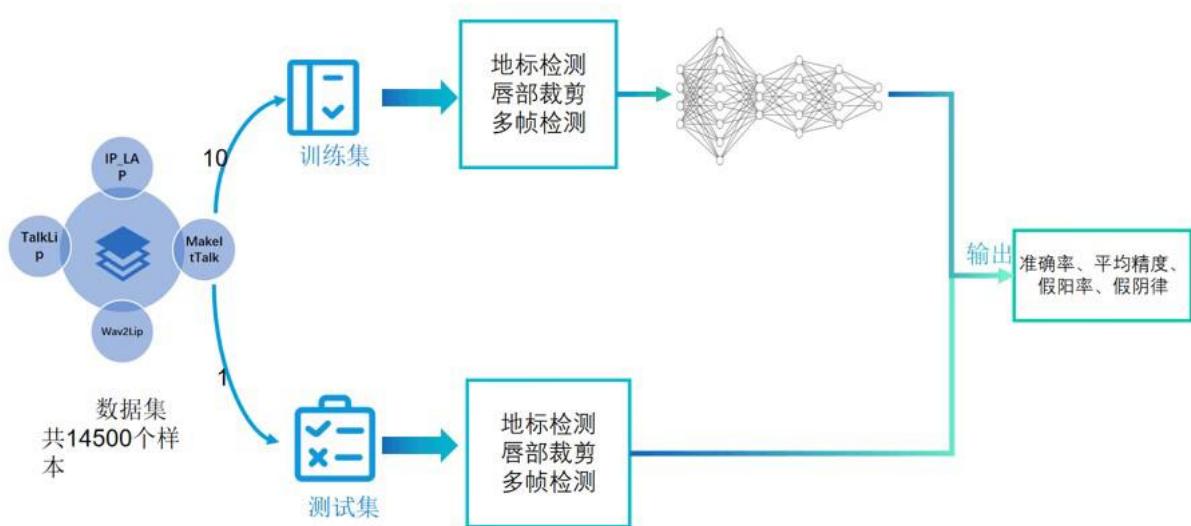


图 5.7 “有效性检验” 过程

我们设置了以下指标对模型性能进行量化考核5.11：

---

表 5.11 准确性、召回率和精度指标

---

指标名称	定义
准确率 (ACC)	用于衡量模型整体预测正确的比例。
平均精度 (AP)	用于衡量模型的总体表现，包括精度和召回率的平衡。
假阴率 (FNR)	正样本中被错误预测为负的比率。
假阳率 (FPR)	负样本中被错误预测为正的比率。

---

测试结果如下5.12：

表 5.12 有效性检验结果对比

模型	AVLip			
	ACC↑	AP↑	FPR↓	FNR↓
CViT	65.54	56.68	0.07	0.61
DoubleStream	75.52	67.72	0.13	0.36
UniversalDetect	50.03	50.02	0.99	<b>0.01</b>
SelfBlended	49.99	52.13	0.07	0.51
RealForensics	<u>91.78</u>	<u>90.14</u>	<b>0.02</b>	0.14
LipForensics	86.13	81.56	0.18	0.10
LipFD	<b>95.27</b>	<b>93.08</b>	<u>0.04</u>	<u>0.04</u>

模型	FF++
----	------

	ACC↑	AP↑	FPR↓	FNR↓
CViT	62.86	54.17	0.24	0.50
DoubleStream	91.02	<b>87.64</b>	<b>0.03</b>	0.14
UniversalDetect	50.43	50.16	0.99	<b>0.01</b>
SelfBlended	64.59	57.93	0.17	0.53
RealForensics	93.57	<u>91.32</u>	<b>0.03</b>	0.10
LipForensics	<u>94.03</u>	<b>93.25</b>	<u>0.04</u>	0.08
LipFD	<b>95.10</b>	76.98	0.06	<u>0.05</u>
模型	DFDC			
	ACC↑	AP↑	FPR↓	FNR↓
CViT	70.99	58.06	<u>0.06</u>	0.50
DoubleStream	77.39	69.28	0.21	0.24
UniversalDetect	49.86	49.94	0.98	<b>0.01</b>
SelfBlended	48.47	49.06	0.15	0.50
RealForensics	<u>92.54</u>	<b>91.62</b>	<b>0.01</b>	0.14
LipForensics	90.75	<u>87.32</u>	0.08	0.11
LipFD	<b>94.53</b>	78.61	0.08	<u>0.04</u>

其中，上述表加粗的数值强调各模型中检测效果最好的模型的准确率，有下划线的数值表示检测效果其次的模型的准确率。我们将整个检验过程中 ACC 阈值设定为 0.5。

通过上述结果可以看出，LipFD 模型在面对三种数据集的测试下，都表现出很优秀的检测率，而且误报率低。团队将 LipFD 模型和 RealForensics 模型的结果进行对比，精

确度 ACC 确实做到了一定程度上的提升。然而，在合作创建的 AVLips 数据集上，LipFD 模型的 AP 值更高，而在 DFDC、FF++ 数据集上的 AP 值出现明显的降低，其中 DFDC 数据集上低 14.34%，在 FF++ 数据集上低 13.01%。

这一结果并不表示 LipFD 检测具有一个低灵敏度，而是因为 LipFD 模型主要针对于 LipSync 技术伪造领域的检测，在利用 LipSync 领域最新的伪造技术进行生成的数据测试集 AVLips 上保持超高的灵敏度，而 DFDC、FF++ 数据集中的数据样本并未过多涉及到 LipSync 领域的技术伪造，所以 LipSync 模型在这两个数据集上的检测灵敏度低于 RealForensics。总而言之，LipFD 模型在准确率 ACC 上在三个数据集上的检验测试都保持了优秀的性能。

### · 鲁棒性测试

在实际遭遇诈骗过程中，对方可能会在换脸之后进行对图像的扰动攻击，且视音频在信道传输过程中难免会出现质量损失。为测试模型在鲁棒性测试下是否依然保持优秀性能。因此，我们在正常换脸后挑选出部分样本视频实施 6 种被广泛使用的扰动攻击，再根据 ROC-AUC 指标参数来检验模型的鲁棒性。扰动攻击如下图 5.8：

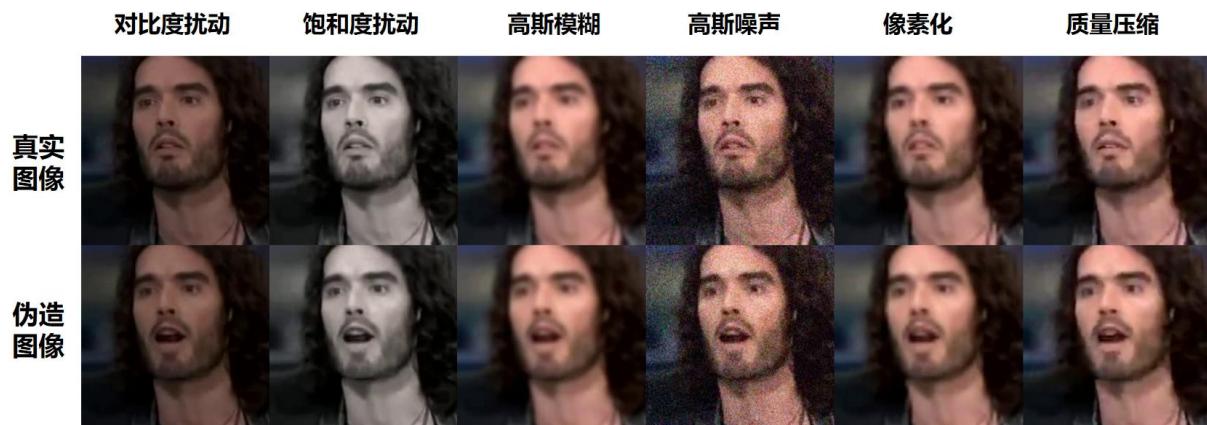


图 5.8 扰动攻击类型

其中，扰动策略参考了图像质量评定标准 (IQA) 领域所定义的 6 中关键性失真类型，即 1) 对比度扰动，2) 饱和度扰动，3) 高斯模糊，4) 高斯噪声，5) 像素化，6) 视频质量压缩。

团队在进行以上六种类型的扰动方法时，还对每种定义了 5 个扰动强度，以更贴合实际地还原实际通话中的各种复杂情况。

表 5.13 针对鲁棒性检验的扰动方法

扰动方法	主要参数	扰动等级				
		1	2	3	4	5
对比扰动	对比增强系数	0.85	0.725	0.6	0.475	0.35
饱和扰动	YCbCr 通道值	0.4	0.3	0.2	0.1	0.0
高斯模糊	高斯核大小	7	9	13	17	21
高斯噪声	噪声方差	0.001	0.002	0.005	0.01	0.05
像素化	像素等级	2	3	4	5	6
压缩	压缩率系数	30	32	35	38	40

最终，测试整体方案如图5.9：

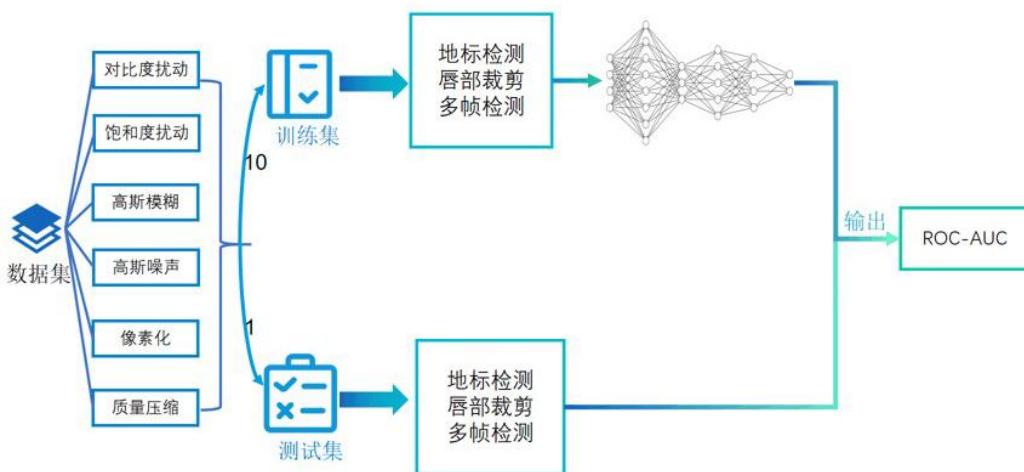


图 5.9 LipFD 伪造检测器鲁棒性测试方案

我们设置了以下指标对模型性能进行量化考核5.14：

表 5.14 准确性、召回率和精度指标

指标名称	定义
ROC 曲线	通过绘制真阳性率 (True Positive Rate, TPR) 与假阳性率 (False Positive Rate, FPR) 之间的关系来评估分类模型的性能。

接下页

表 5.14 – 续

指标名称	定义
AUC	ROC 曲线下与坐标轴围成的面积。

检测结果见图 5.10：

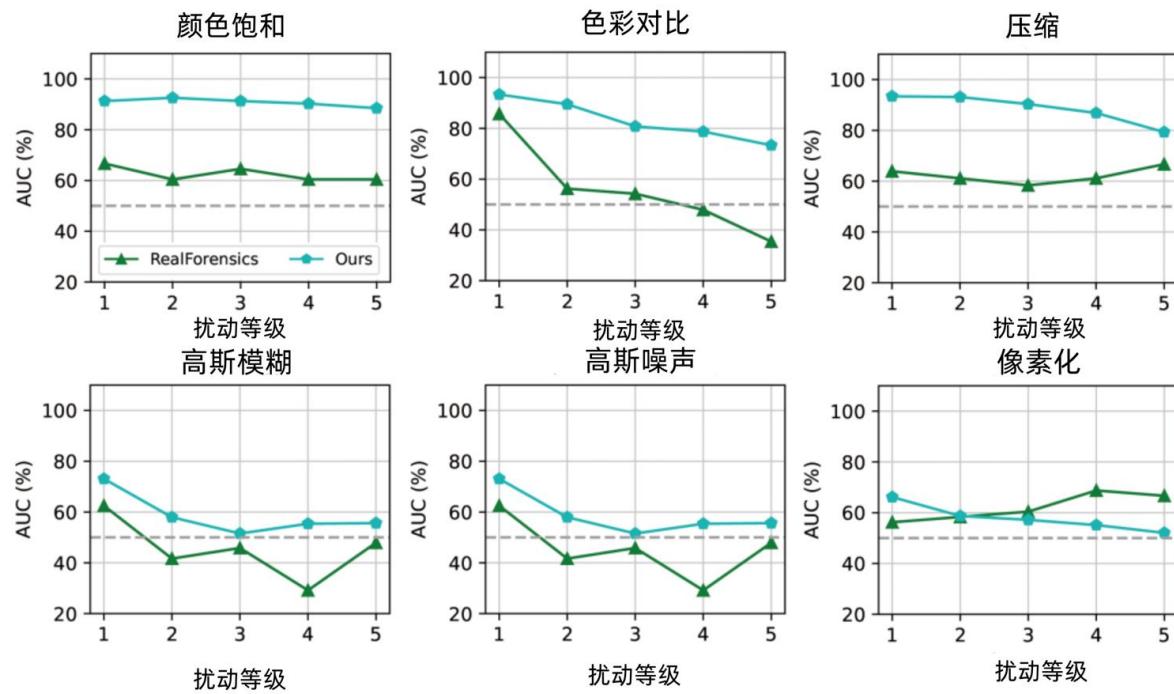


图 5.10 LipFD 伪造检测器鲁棒性测试结果

在 HLS 空间内，我们的检测模型对基于线性变化的颜色饱和、色彩扰动和视频压缩都有着优秀的抗干扰能力，在 5 个扰动攻击等级下，都保持了相较于 RealForensics 模型更好的稳定性。

在高斯模糊检验测试中，算法固定高斯核的大小，调整标准偏差以此来还原模拟不同程度下的图像模糊，而测试结果表明高斯模糊和像素画技术都会对我们的检测器的检测效能产生一定的不良影响。团队调研资料后，讨论分析，可能与高频信号对模型的干扰性有关。因此，团队在实际应用中也是采用了模糊后的数据集进行微调训练，以得到更好的模型检测性能。

## (2) 音频伪造

我们的音频伪造检测技术别技术测试将在有效性测试和鲁棒性测试下分别进行测试。

### · 有效性测试

我们着重关注模型的精确度。为此，我们将真实样本与伪造样本同时输入检测器，并利用模型进行分析与判断，生成识别结果。这些结果会与预先准备的标签集进行比对，以验证其准确性。为了科学、客观地评估检测器的性能，我们还与其他两种模型进行对比，并采用了等错误率（EER）作为评分标准，通过对真实错误率与假正错误率的平衡点，提供了一个量化且直观的衡量指标。

实验中，本团队使用 RawNet2[90] 作为基线，Adaptive Moment Estimation (Adam) 作为优化器，学习率为 0.0001，批量大小为 32，损失权重  $\lambda$  设为 0.5。

为更好地展现我们模型的效果，我们获取了近年来其它音频伪造检测方面有突出成果的模型的训练成绩并进行对比。我们挑选出 LFCC-LCNN[91]、原始 RawNet2、WavLM[92] 和 Wav2Vec2-XLS-R[93] 为基线模型进行对比试验。LFCC-LCNN 是将 LFCC[94] 特征提取与 LCNN[95] 分类器 (DNN) 相结合的方法，在 ASVspoof 2021 Speech DeepFake 赛道中取得了第二名的成绩；原始 RawNet2 模型基于 DNN 说话人嵌入提取，以原始波形作为输入。该模型使用一种名为特征图缩放的技术，该技术可以像 squeezeexcitation[96] 一样缩放特征图。它在 ASVspoof 2021 Speech DeepFake 赛道中表现最佳；微软的 WavLM 模型是一种自监督的预训练多语言模型，可用于各种下游语音任务；Wav2Vec2-XLS-R 模型是 Meta AI 开发的用于语音相关任务的大规模多语言预训练模型。

基于此，我们的测试过程如图 5.11：

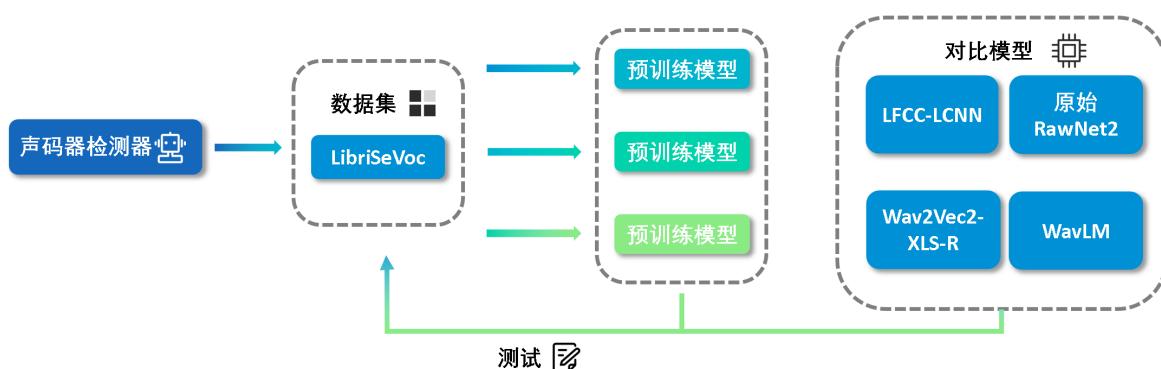


图 5.11 声码器检测器测试流程

测试使用的原始 LibriSeVoc 数据集包含的声码器数量、频率、训练集大小、验证集大小和测试集大小见表 5.15：

---

表 5.15 测试使用的数据集信息

Dataset	#Vocoder type	Frequency	Training size	Dev size	Testing size
LibriSeVoc	6	24kHz	55,440	18,480	18,487

数据集由真实样本和伪造样本构成。一个伪造的例子如图5.12：

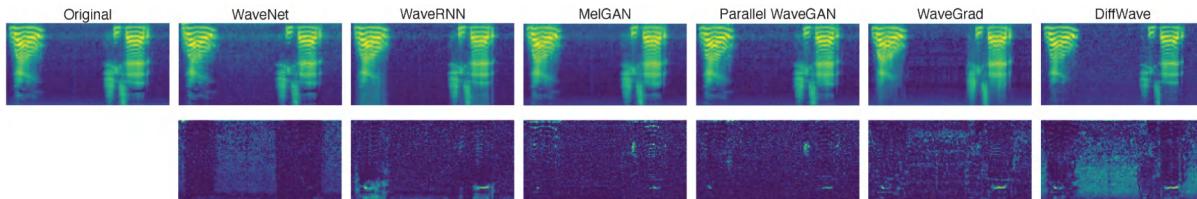


图 5.12 待检测音频的梅尔图谱：左一为原始样本，右六为六种声码器合成的伪造样本，下方为声码器对原音频造成的差异

我们使用的评估性能如下表5.16。其中，ACC，FNR 和 FPR 指标能够直观地反映模型在所有样本上的综合表现，是常用的评估指标；EER 是一个综合性指标，能够平衡假阴率和假阳率，提供一个量化且直观的衡量标准，特别适用于需要平衡两类错误的场景。

表 5.16 评估指标

指标名称	定义
假阴率 (FNR)	正样本中被错误预测为负的比率。
假阳率 (FPR)	负样本中被错误预测为正的比率。
等错误率 (EER)	假阴率和假阳率相等时的错误率。

实验结果如表5.17。前三列的结果表明，我们的模型在 LibriSeVoc 上实现了最低 EER=0.13%，在 WaveFake 上实现了最低 EER=0.19%，明显优于在每个数据集上重新训练的其他基线。

然后，我们在带有欺骗和虚假音频的 ASVspoof 2019 数据集上评估我们的方法。最

---

后一列的结果显示在 ASVspoof 2019 上实现了最低 EER 4.54%，即我们的模型在检测各种音频欺骗攻击方面表现最佳。

表 5.17 对比其他方法的测试结果

Methods	LibriSeVoc	WaveFake	ASVspoof
LFCC-LCNN	0.14	<b>0.19</b>	11.60
RawNet2	0.17	0.32	6.10
WavLM	0.45	2.92	6.94
Wav2Vec2-XLS-R	1.54	2.33	13.48
<b>Ours</b>	<b>0.13</b>	<b>0.19</b>	<b>4.54</b>

通过上述结果可以看出，我们的模型在面对三种数据集的测试下，都表现出很优秀的检测率，而且误报率低。具体来说，在 LibriSeVoc 数据集上，我们的模型的错误率为 0.13%，在 WaveFake 数据集上为 0.19%，在 ASVspoof 数据集上为 4.54%，均优于其他方法。

相比之下，LFCC-LCNN 在 LibriSeVoc 和 WaveFake 数据集上的表现也较为出色，但在 ASVspoof 数据集上的错误率较高，为 11.60%。RawNet2 在 ASVspoof 数据集上的表现较好，错误率为 6.10%，但在其他两个数据集上的表现不如我们的模型。WavLM 和 Wav2Vec2-XLS-R 在所有数据集上的表现均不如我们的模型，错误率较高。

总而言之，我们的模型在三个数据集上的测试结果均表现出色，展示了其在不同数据集上的强大检测能力和鲁棒性。特别是在 ASVspoof 数据集上，我们的模型的错误率显著低于其他方法，表明其在处理复杂数据时具有较高的准确性和稳定性。可以看到，本团队使用的模型在各个数据集上都呈现出了较好的结果，拿到了最低的 EER。这表明我们的模型在有效性测试下有极高的准确度。

### · 鲁棒性测试

我们使用添加了扰动后的重构 LibriSeVoc 数据集进行测试，重构数据集模拟了采样率改变和噪声干扰两种常见的干扰环境，有助于我们更好的理解模型在拟实际环境下的表现。通过在重构后的数据集上进行测试，我们可以获取模型在不同场景下的性能，

#### 关于采样率的扰动

在实际场景中，用户使用的手机或电脑等电子设备型号不同，而不同的录音设备可能有不同的采样率限制。为保证在不同采样率下该模型都能稳定检测输出成果，我们针对模型在不同采样率下的鲁棒性做了测试。测试过程中，首先我们将输入语音重新采样

为中间采样率（8kHz、16kHz、22.05kHz、32kHz 和 44.1kHz），然后重新采样回原始采样率（24 kHz）。一个例子如图 5.13：

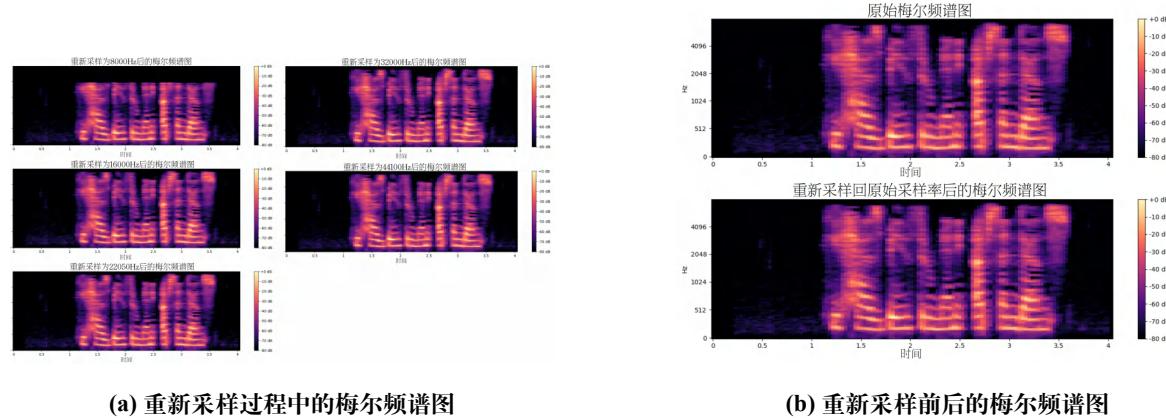


图 5.13 模拟干扰采样率测试范例

### 关于噪声的扰动

此外，除了联系人对话的声音，还可能有各种环境音的干扰，为保证在不同通话场景下我们的模型均能稳定输出，我们对模型在有噪声干扰下的鲁棒性做了测试。测试过程中，我们通过添加三个 SNR 值（8dB、10dB 和 20dB）的单个预录人群噪音样本来引入背景噪音。我们随机选择原始、重新采样或嘈杂的语音片段，概率分别为 40%、40% 和 20%。一个例子如图 5.14：

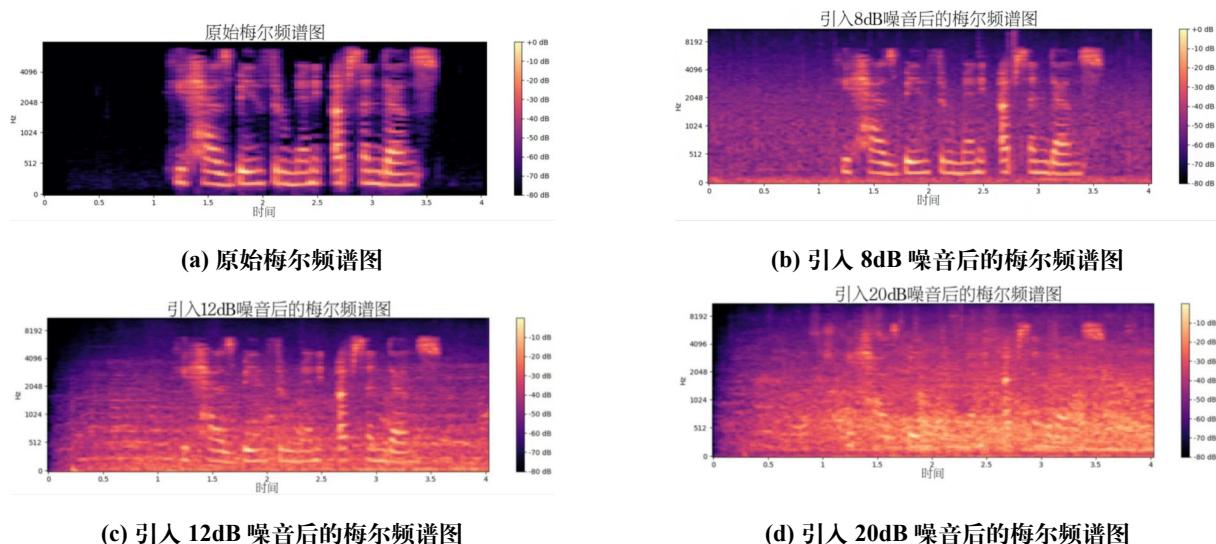
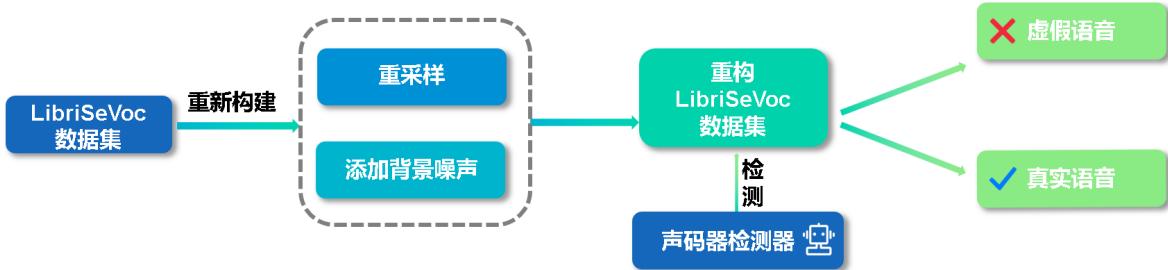


图 5.14 模拟噪声数据集测试范例

基于以上，我们的测试过程如图 5.15



我们仍然使用 ACC, FNR, FPR, EER 指标进行评估, 解释见前文 5.1.3。鲁棒性测试下实验结果见图 5.16。

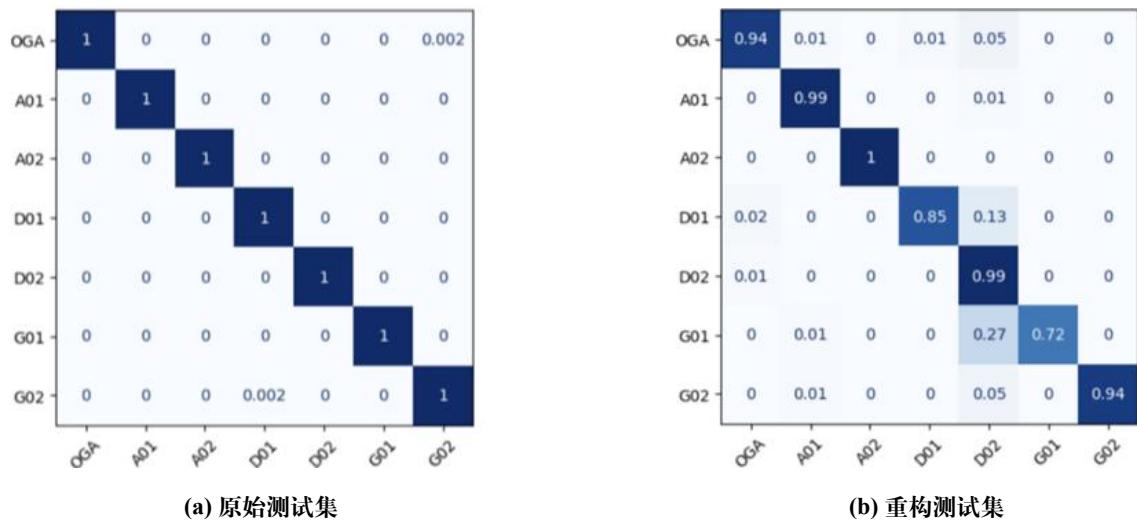


图 5.16 在 LibriSeVoc 上评估的混淆矩阵

(OGA: ground truth, A01: WaveNet, A02: WaveRNN, D01: WaveGrad, D02: Diffwave, G01: MelGAN, G02: Parallel WaveGAN.)

其中的混淆矩阵进一步比较了原始 LibriSeVoc 集 (检测 EER 为 0.13%) 和添加扰动后处理数据集 (EER 为 2.73%) 上的检测和声码器识别性能。可以看到, 在重构测试集上高分仍然集中于对角线, 最高为 1, 最低为 0.72。这表明我们的方法可以提取用于声码器识别的判别性声码器级特征, 并且对常见的数据后处理操作也具有鲁棒性。

## 5.2 系统功能测试

为了测试我们系统在各个场景下的识别诈骗功能是否正常, 我们将模拟在各种诈骗环境下, 使用各种诈骗手段对系统进行诈骗攻击。

我们将诈骗攻击手段分为技术伪造流、事实引导流、伪造流和引导流的结合。我们选择在此基础上生成 6 个在日常生活最可能发生的诈骗攻击实例，它们具体的攻击方式如下表（✓ 表示含有此诈骗手段）。1 号攻击为我们设置的对照组，用来测试我们的系统会不会对非诈骗通话产生误判。人脸和音频结合的伪造攻击是为了模拟视频通话的诈骗，单一的音频通话伪造是模拟音频通话的诈骗，而引导式诈骗是模拟 AI 文本生成和传统的诱导式话术哄骗，如图 5.17。

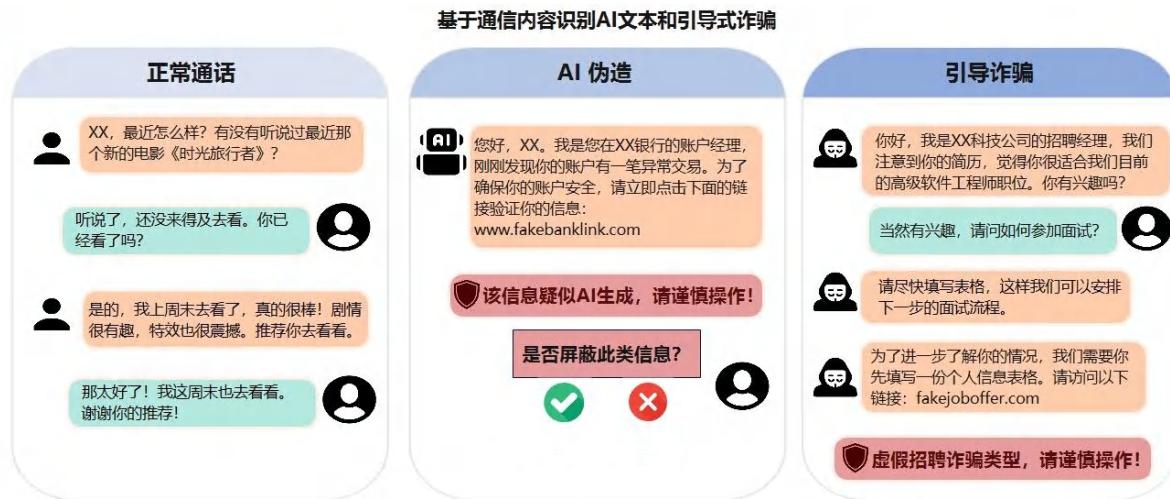


图 5.17 测试任务样例

我们使用的攻击类型如表 5.18：

表 5.18 诈骗攻击实例

攻击序号	人脸伪造	音频伪造	引导式诈骗
1	-	-	-
2	-	-	✓
3	-	✓	-
4	-	✓	✓
5	✓	✓	-
6	✓	✓	✓

我们将生成的攻击在我们模拟的各种环境条件下进行攻击，包括光线是否极端，环境声音是否嘈杂，网络连接是否良好。然后我们统计系统在相应环境下对攻击正确判断的数量。结果如表 5.19：

---

表 5.19 模拟攻击测试结果

环境条件			正确判断数
网络延迟	光线极端	环境声音嘈杂	
-	-	-	6
-	-	✓	6
-	✓	-	5
-	✓	✓	5
✓	-	-	5
✓	-	✓	4
✓	✓	-	4
✓	✓	✓	4

整体的结果显示，我们的系统在检测环境良好时，基本能够正确地判断所有的诈骗攻击真伪。当存在一定的光线极端或者环境噪声时，我们的模型性能会稍微下降，光线昏暗、面部遮挡、环境噪音太大等情况确实会影响系统在某些方面的检测能力。而当网络存在延迟卡顿时，我们系统的检测效果开始出现一定的下滑，原因是网络的延迟卡顿可能导致视频画面的模糊以及声音的不流畅，非常考验系统的判断能力。不过考虑到在这些不良检测环境条件下存在的情况下，诈骗分子的诈骗效果也会大打折扣，我们的模型已经能够实现很好的判断性能了。

### 5.3 系统易用测试

为了评估本系统在实际使用中的用户体验，我们设计了系统易用性测试。我们从易浏览性和易操作性两个方面进行评估，旨在展示系统在部署落地后的真实效果。评估结果如表5.20：

---

表 5.20 易用性测试表

---

测试项目	测试过程描述	测试结果
风格一致性	页面美术风格、菜单、主界面、功能弹窗、字体、列表、数据精度的风格是否一致	系统页面风格一致
易浏览性	输入、输出结构与规整，显示效果简洁直观	系统组件提示操作信息，方便指导用户操作
	输出内容专业、准确且可读性强	分析报告用语简明清晰，便于用户读懂
易操作性	用户自由度高，能够自行选择开启和关闭多种功能	用户可自由选择是否开启检测，随时关闭摄像头，随时获取分析报告，自由选择离线文件上传分析
	软件操作简单，系统支持标准鼠标、键盘操作，支持鼠标单击、双击和右键操作，支持触屏，支持快捷键操作	均配置完全

## 5.4 系统可靠测试

为了评估本系统的可靠性，我们设计了系统可靠性测试。我们从成熟性、容错性、易恢复性和数据校验机制个方面进行全面评估，综合考虑实际应用场景中可能出现的风险和干扰，旨在全方位测试本项目的鲁棒性。测试结果如下5.21：

表 5.21 可靠性测试表

测试项目	测试过程描述	测试结果
成熟性	系统接收文件达到上限时，系统不崩溃，不出现错误数据，不丢失数据	系统在接收文件达到上限时给出预警
	录入错误信息时，系统不崩溃，数据不丢失	系统给出相应的提示信息
容错性	用户误操作时，系统能做出相应屏蔽以维持功能稳定	用户操作不合规时，系统撤销相应操作，撤销后功能正常
	对各种误操作有对应提示	对用户的各种误操作，系统能给出相应提示。如用户上传了错误格式的文件，输入了不符合规范的密码，系统会显示相应提示信息
易恢复性	系统因不可抗力崩溃前，能即时保存重要信息，崩溃后，能快速重建系统	系统数据库有相应保障
数据校验机制	对各模态输入数据之间的逻辑关系进行校验，保证各类数据分类存放	系统符合该项操作，如能正确处理 MP4 和 wav 混合文件
	对各维度输出的逻辑关系进行校验，返回合理结果	系统正确处理多维输出，综合分析后返回结果

---

## 第6章 作品总结

### 6.1 作品特色与创新点

“一盾当关”系统面向当前诈骗防御存在的痛点与缺陷，采用多模态多角度的方式全面地判断诈骗并及时预警用户，主要分为面向多样化诈骗手段的及时预警、针对先进诈骗手段的精准突破、面向通信内容、载体和行为的风险解释、面向风险形象的大数据推送和面向应用的易部署性和易拓展性等系统特色。

#### 6.1.1 推出全面且高效的多模态风险内容识别系统

目前的防诈系统在实际场景中大多没有很好的实用性和易操作性，无法真正意义上帮助用户很好地规避诈骗带来的风险和损失。而本防诈系统，具有高度全面性、易操作性。系统不仅可以全方位地检测技术伪造痕迹，还可以检测基于技术事实的引导式诈骗话术。用户在实时通话中，无需任何额外操作，系统会自动检测对方的诈骗痕迹，并提供风险摘要内容。此外，为进一步方便用户，系统还提供离线文件检测功能，可以检测离线视频、音频、图片、文本。

#### 6.1.2 基于跨模态时序不一致性的人脸伪造检测算法

随着 LipSync 伪造方法的不断演进，伪造人脸高度逼真，视觉上的伪影十分隐蔽、不可察觉，导致现有检测器在应对先进的 LipSync 方法时效果欠佳。但是，我们发现伪造内容中的视频帧和音频帧会存在细微的不一致性，基于此，**团队联合导师提出了一种针对唇同步伪造的检测方法 LipFD**，关注帧与帧的联系和连续帧时序对齐，深入探究唇部运动和音频信号之间固有的内在差异，并针对伪造视频中普遍存在的语音连续性和视频离散性不匹配的问题来捕捉细微的音视频时序和空域伪影，最终实现有效区分真实与伪造。

#### 6.1.3 基于交叉模态情绪一致性的诈骗心理识别方法

诈骗分子在诈骗过程进行中不可避免地会心虚或紧张，并不自觉地表露在面部表情、语调和“小动作”等情绪细节中。本系统能够通过视频中的面部表情和音频分析来初步评估通话人与诈骗相关的情绪，同时捕捉视频中通话人的异常动作。这种多重分析方法通过重合情绪分析结果而精准定位异常情绪时间点，并与相应时间段的文本内容进行对照分析。通过这种综合分析，系统可以有效且准确地推测出检测对象的诈骗动向和诈骗心理，提升了诈骗检材的鉴定精度与可靠性。

#### **6.1.4 高度可解释的诈骗风险检测解释方案**

技术的迅猛进步使得诈骗手段变得越来越复杂和隐蔽，带来了新的挑战。面对多样且复杂的诈骗形式，传统的检测系统显得力不从心，并且难以全面分析和解释诈骗风险的根源，无法得到用户的信任。因此，我们团队研发了一种多模态的综合判断机制。该系统不仅深入分析通话内容的诱导性、通话载体的真伪和通话行为中的情绪细节，还创新性地整合了动作特征、面部表情识别和语音情绪分析等多个维度，实现了对诈骗分子心理活动的跨模态、多层次剖析。这一创新方法使系统能够精准捕捉诈骗行为的蛛丝马迹，能够定位诈骗发生的时间点。并基于详实的数据生成一份全面、透彻且理由充分的诈骗风险解释报告，从而显著提升用户对系统决策过程的理解与信任。此外，我们还接入 LRMS 大推理模型，向其输入风险摘要报告和整个通话数据，借助其思维链理解能力，用户可以更好地通过询问该 ai 助手来全面清晰地了解整个通话过程中的相关风险点，提高可行度和理解力。

。

#### **6.1.5 面向风险形象的大数据推送**

为了向用户提供更加个性化的服务，系统将基于用户的注册身份，独立管理其相关数据信息，实现便捷的登录和使用体验。同时，系统将记录和分析用户的历史通讯数据，通过对用户通话内容的深入挖掘，识别出用户易遭受的诈骗类型，以及其在心理防御方面的薄弱点，从而精确地勾勒出用户的受骗风险画像。结合大数据技术，系统将基于“用户画像”精准定位防御弱点，持续推送与之对应的诈骗案例和教育资料，帮助用户识别潜在风险。此外，系统不仅限于已发生的诈骗场景，还会主动推送用户未曾接触过的诈骗手法，拓展用户的风险认知面，做到全方位、多角度地提升用户的防诈骗意识。通过不断强化用户的反诈认知和心理防线，系统将有效帮助用户构建全面的自我保护机制，真正做到防患于未然。

#### **6.1.6 面向应用的易部署性、易拓展性、易维护性**

目前，诈骗已经从传统的通讯诈骗上升至整个社交平台上的诈骗，同时诈骗的手段方式也日益复杂多样，诈骗检测系统需要拓展出新的内容功能才能应对未来更具有挑战性的诈骗手法。但大多数现有的系统架构复杂，使得企业在特定需求场景下部署系统时要进行大规模的调整，部署和维护的成本极高；而且系统模块间复杂的数据流交互，系统对模块复杂的调用过程使得系统拓展新功能的成本极高。“一盾当关”系统采用模块化设计和标准化接口，实现了跨平台兼容。系统采用模块化设计、前后端分离架构，各

---

个功能模块相对独立，可以根据实际需要进行组合和配置，实现各种应用场景下的灵活部署；同时，各个功能模块可以独立扩展和升级，方便技术更新迭代和系统维护。

## 6.2 应用推广

本项目致力于构建一个融合引导式诈骗识别与基于内容生成的伪造式诈骗的实时诈骗检测平台，以应对当前诈骗手段日益多样化及高度技术化的挑战。长期以来，诈骗、谣言等多模态风险内容对民众财产安全构成了严重威胁，给我国造成了巨大的经济损失。我国已投入大量资源致力于打击电信诈骗、舆情监管，成效显著，诈骗案件数量显著下降。然而，面对不断演变的新形势，特别是多模态引导式诈骗与基于内容生成技术的伪造诈骗的兴起，现有的检测手段显得力不从心，市场上缺乏广泛集成与高效应用的解决方案。

针对这一痛点，我们的系统创新性地基于多模态的基础架构，深入洞察并覆盖了现代诈骗所普遍采用的各类技术手段，并能实时高效地进行监控防护。这一设计旨在打破传统检测方法的局限，期望系统能够在社交网络、金融、司法、电商与市场营销等各领域的应用场景中展现出卓越的性能，从而帮助广大用户有效规避诈骗风险，无论面对何种形式的诈骗手段都能保持高度的警觉与防范能力。我们坚信，通过这一努力，将为社会带来更加安全、可靠的防诈骗解决方案，为守护公众财产安全贡献积极力量。

### 6.2.1 社交网络

随着网络媒体和社交通讯软件的普及，人们的生活日益数字化、网络化。这虽然极大地丰富了我们的日常交流和信息获取方式，但也在无形中为诈骗分子提供了更多可乘之机。诈骗分子不再局限于信件、电话诈骗，而是利用这些媒体平台实施面向个体的更高级、更难以察觉的诈骗行为。此外，基于伪造式内容而捏造出来的“谣言”面向公众，大面积地传播虚假的误导性信息。而社交网络领域在防御诈骗、谣言的呢过多模态风险内容时却缺乏统一的风险过滤模块。

#### (1) 谣言信息

在诸如微博等社交平台上，诈骗分子通过 deepfake、语音合成等技术来生成伪造视频、音频，基于高质量的虚假内容，来捏造并散布谣言、虚假新闻等不实言论，以此给社会带来恐慌，造成十分恶劣的影响。面对此问题，传统的舆情监控机制不能做到及时反应，遏制谣言传播，无法最大程度降低损失。而本项目能够实时监控平台上散播的内容，并精准识别其中是否存在伪造内容，以第一时间向用户发出警告，防患于未然。

#### (2) 即时通讯

---

在如微信等即时通讯平台上，诈骗分子通过伪造好友账号，利用视频通话功能展示事先录制的“真人”视频或利用实时伪造手段生成伪造图像，结合精心设计的诈骗话术，诱导受害者转账或提供个人信息。面对此问题，社交网络普遍缺乏即时、高效的诈骗检测机制。传统方法难以捕捉非语言信号，而本项目能够全面分析语音、视频、文本等多模态信息，识别诈骗行为中的微妙迹象。而本项目通过“实时多模态情感与行为分析系统”，结合用户行为模式、情感反应及语言特征，实现对通话过程的实时监控与诈骗预警，有效阻断诈骗行为。

### **(3) 视频网站**

在如抖音等视频网站平台上，诈骗分子发布伪造的名人代言视频，利用深度伪造技术合成名人脸部，并配以误导性广告词，诱导用户购买假冒伪劣产品。传统方法依赖用户举报和人工审核，反应滞后，难以应对快速演变的诈骗手段。视频网站面临海量内容审核难题，传统方法难以快速甄别。同样的，本项目可以对视频进行自动检测，为视频审核工作减轻压力。

## **6.2.2 司法取证**

在诈骗案件中，犯罪分子可能通过多种手段造假通话录音和视频证据，企图逃脱法律制裁。传统取证手段难以完全确保辨别出证据真伪，且无法自动化高效地发现犯罪分子由于紧张害怕而表现出来的情绪、话语、动作的失配性，导致案件审理进展缓慢。面对此类问题，本项目设计的先进检测分析系统，可以对相关内容信息进行处理，从多个维度综合分析，精准识别伪造痕迹，并且提供自动化的技术手段来捕捉犯罪分子异常心理情绪，期望能对司法鉴定工作提供一定帮助。

## **6.2.3 金融领域**

金融领域面临的诈骗风险通常来自涉及金钱转账的人脸识别或身份认证欺诈。金融机构和个人都有可能受到这种诈骗的攻击。

### **(1) 机构**

在保险机构中，保险理赔需要能表示事实性的资料证明。诈骗分子可能会通过人脸替换、声音合成来伪造事故、伤害或者疾病，以诱导式话术来骗取保险金。本系统可以识别诈骗分子话术中的哄骗引导部分和内容伪造部分，警告用户存在被诈骗风险。

### **(2) 个人**

在银行机构中，办理金融业务需要申请人的详细身份信息。诈骗分子可能会通过技术手段冒充银行或其他可信赖的金融机构员工，以此来骗取受害者的个人信息或者财

---

产。本项目可以检测对方是否存在人脸、语音的伪造可能，以及对方是否使用诱导性话术哄骗用户透露个人信息，从而揭穿诈骗分子的谎言。

#### 6.2.4 电商与市场营销

电商与市场营销主要面临着线上交易和广告宣传的诈骗风险。

##### (1) 线上交易平台

在如“京东商城”等线上交易平台中，用户和商家可以进行实时交互。然而，诈骗分子冒充正规商家实施诈骗，利用伪造技术隐藏真实身份，用户难以察觉，甚至难以维权。针对此风险，本系统研发的多模态诈骗检测和识别系统，可以实时监控用户和商家的聊天记录，分析并检测音视频内容，及时返回检测结果，帮助消费者免受诈骗行为，提供安全保障。

##### (2) 广告宣传

在如“巨量引擎”的广告宣传平台中，商家制作广告内容并上传展示。然而，诈骗分子利用伪造技术伪造广告内容，误导消费者，导致用户无法分辨真假广告，进而受到欺骗。针对这一问题，本系统开发了高精度伪造内容检测技术和基于事实的引导式诈骗检测技术，能够有效检测伪造的音视频广告内容，识别诱导性广告。系统保证检测过程高效，为用户提供可信赖的广告内容鉴别服务，帮助他们识别并避开虚假和诈骗广告，确保用户权益不受侵害。

### 6.3 作品展望

在综合了实际场景测试反馈及内部讨论后，本团队已制定出后续升级和应用落地的详细规划，以进一步提升系统的性能和功能，并对系统进行工业化全面推广：

**(1) 实时性能优化：**鉴于诈骗行为的突发性和对连续实时监控的需求，我们的系统已能处理每秒 30 帧的视频数据。未来，我们计划优化网络结构，采用多线程并发处理，并引入边缘-云端协同架构，在 5G 网络下实现低于 50ms 的诈骗风险拦截，覆盖直播电商、远程医疗等高实时性场景，提升系统响应速度与处理效率。同时，我们将尝试集成多模态预训练大模型，实现跨文本、音视频、行为数据的统一表征学习，解决“伪造技术超前于检测技术”的行业难题。我们还将向公安部门申请数据与试点，结合实际反馈不断优化性能，确保监控的连续性和实效性。

**(2) 加强数据处理和分析：**随着团队技术的成熟和应用市场的拓展，我们将迭代开发更具性价比的系统，能消耗更低的算力资源，兼容更低性能硬件设备；我们将基于大模型技术、伪造技术、诈骗手段的发展，不断更新我们的检测技术，并融合更为有效的

---

大模型；计划引入先进的机器学习算法来深入分析已有的诈骗案例数据，并在通话过程中能基于历史数据向用户提供解释性预警和实例，让用户能提前有所防范。

**(3) 功能扩展：**展望未来，我们将利用积累的经验和资源大力投入技术研发，推动持续的技术创新。我们计划将“一盾当关”系统发展为一款技术强大、功能完善、性能高效的系统插件，应用于数字通讯、网络安全、社交平台等所有互联网领域，拓展功能包括但不限于异常行为监控、身份验证、行为心理分析、司法辅助审讯等。这些增加功能将进一步发挥系统的潜力，守护互联网中的每一份人际交流，让彼此的沟通更加安全、顺畅。

**(4) 生态共建：**联合反诈中心、金融机构与科技企业，构建诈骗特征共享联盟链，实现跨平台风险数据的安全流通与联合建模；推出防诈认知教育模块，基于用户历史受骗画像生成个性化反诈训练课程，从技术防御迈向“人机协同防御”。在政策与市场的双重驱动下，本系统有望成为数字社会的“免疫中枢”——既为个体用户筑起实时防护屏障，又为金融、政务、医疗等关键领域提供风险管理基础设施。团队将持续迭代技术纵深与生态连接能力，目标一年内实现“系统落地推广”、三年内“千万终端覆盖”，助力数字信任体系的重构。

## 参考文献

- [1] 中国网络空间研究院, 中国互联网发展报告 2024. 北京: 商务印书馆, 2024.
- [2] P. Institute and I. Security, “2023 cost of a data breach report,” 2023. Accessed: August 2, 2024.
- [3] W. E. Forum, *World Economic Forum Annual Report 2021-2022*. Geneva: World Economic Forum, 2022.
- [4] Y. Fan, M. Xie, P. Wu, and G. Yang, “Real-time deepfake system for live streaming,” in *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pp. 202–205, 2022.
- [5] MyHeritage, “Deep nostalgia.” <https://www.myheritage.com/deep-nostalgia>, 2021.
- [6] W. Zhong, C. Fang, Y. Cai, P. Wei, G. Zhao, L. Lin, and G. Li, “Identity-preserving talking face generation with landmark and appearance priors,” 2023.

- 
- [7] T. Zhu, J. Chen, R. Zhu, and G. Gupta, “Stylegan3: Generative networks for improving the equivariance of translation and rotation,” 2024.
  - [8] J. Guan, Z. Zhang, H. Zhou, T. HU, K. Wang, D. He, H. Feng, J. Liu, E. Ding, Z. Liu, and J. Wang, “Stylesync: High-fidelity generalized and personalized lip sync in style-based generator,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
  - [9] W. Zhang and C. Wang, “Development of real-time rendering technology for high-precision models in autonomous driving,” *arXiv preprint arXiv:2302.00291*, 2023.
  - [10] J. M. Martyn, Y. Liu, Z. E. Chin, and I. L. Chuang, “Efficient fully-coherent quantum signal processing algorithms for real-time dynamics simulation,” *The Journal of Chemical Physics*, vol. 158, no. 2, p. 024106, 2023.
  - [11] V. Amaral, S. R. Lima, T. Mota, and P. Chainho, “Exploring webrtc technology for enhanced real-time services,” in *New Perspectives in Information Systems and Technologies, Volume 2*, pp. 43–52, Springer, 2014.
  - [12] A. Saha, W. Hamidouche, M. Chavarriás, F. Pescador, and I. Farhat, “Performance analysis of optimized versatile video coding software decoders on embedded platforms,” *Journal of Real-Time Image Processing*, vol. 20, no. 120, 2023.
  - [13] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” 2016.
  - [14] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, “Deep voice 3: Scaling text-to-speech with convolutional sequence learning,” 2018.
  - [15] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” *arXiv preprint arXiv:1806.04558*, 2018.
  - [16] E. Casanova, J. Weber, C. Shulby, A. Candido Junior, E. Gölge, and M. A. Ponti, “Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone,” *arXiv preprint arXiv:2112.02418*, 2021.

- 
- [17] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” *arXiv preprint arXiv:2010.05646*, 2020.
  - [18] K. Shen, Z. Ju, X. Tan, Y. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian, “Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers,” *arXiv preprint arXiv:2304.09116*, 2023.
  - [19] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, “Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings,” *arXiv preprint arXiv:1910.10838*, 2019.
  - [20] Y. A. Li, C. Han, V. S. Raghavan, G. Mischler, and N. Mesgarani, “Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models,” *arXiv preprint arXiv:2306.07691*, 2023.
  - [21] OpenAI, “Introducing chatgpt.” <https://openai.com/blog/chatgpt/>, 2022.
  - [22] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
  - [23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.
  - [24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
  - [25] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” *arXiv preprint arXiv:1801.06146*, 2018.
  - [26] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
  - [27] J. Yang, C. Zhang, *et al.*, “Multi-modal emotion recognition with transformer-based models,” *IEEE Transactions on Affective Computing*, May 2020.

- 
- [28] M. Pantic and L. J. Rothkrantz, “Automatic analysis of facial expressions: The state of the art,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.
  - [29] A. Graves, A.-r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6645–6649, IEEE, 2013.
  - [30] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
  - [31] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
  - [32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
  - [33] T. Soukupova and J. Cech, “Real-time eye blink detection using facial landmarks,” in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9, IEEE, 2016.
  - [34] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “Mesonet: a compact facial video forgery detection network,” *arXiv preprint arXiv:1809.00888*, 2018.
  - [35] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics++: Learning to detect manipulated facial images,” *arXiv preprint arXiv:1901.08971*, 2019.
  - [36] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, “Protecting world leaders against deep fakes,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 38–45, 2019.
  - [37] Y. Li *et al.*, “Joint audio-visual deepfake detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
  - [38] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, “Lips don’t lie: A generalisable and robust approach to face forgery detection,” *arXiv preprint arXiv:2012.07657*, 2020.

- 
- [39] C. Feng, Z. Chen, and A. Owens, “Self-supervised video forensics by audio-visual anomaly detection,” *arXiv preprint arXiv:2301.01767*, 2023.
  - [40] J. L. Elman, “Finding structure in time,” *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.
  - [41] J. B. Watson, *Behaviorism*. New York: W.W. Norton & Company, Inc., 1930.
  - [42] I. Sysoev, “Nginx.” <https://nginx.org/>, 2004.
  - [43] uWSGI Team, “uwsgi.” <https://uwsgi-docs.readthedocs.io/>.
  - [44] L. Weifeng, S. Tianyi, L. Jiawei, L. Boheng, Y. Dongyu, L. Ziyou, and W. Run, “Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes,” in *Neural Information Processing Systems*, Springer, 2025.
  - [45] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
  - [46] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.
  - [47] B. Heo, S. Chun, S. J. Oh, D. Han, S. Yun, J. Choe, and Y. Yoo, “Adamp: Slowing down the weight norm increase in momentum-based optimizers,” *arXiv preprint arXiv:2006.08217*, 2020.
  - [48] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.
  - [49] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

- 
- [50] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” *arXiv preprint arXiv:2312.00752*, 2023.
- [51] H. Ma, S. Lei, T. Celik, and H.-C. Li, “Fer-yolo-mamba: Facial expression detection and classification based on selective state space,” *arXiv preprint arXiv:2405.01828*, 2024.
- [52] C. Xue, W. Zhang, Y. Hao, S. Lu, P. Torr, and S. Bai, “Language matters: A weakly supervised vision-language pre-training approach for scene text detection and spotting,” *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.
- [53] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “Yolov4: Optimal speed and accuracy of object detection,” 2020.
- [54] J. Yang, Z. Liu, B. Jin, J. Lian, D. Lian, A. Soni, E. Y. Kang, Y. Wang, G. Sun, and X. Xie, “Hybrid encoder: Towards efficient and precise native ads recommendation via hybrid transformer encoding networks,” *arXiv preprint arXiv:2104.10925*, 2021.
- [55] A. Ball and M. Duke, “How to cite datasets and link to publications.” <http://www.dcc.ac.uk/resources/how-guides>, 2015.
- [56] G. Dellaferreira, F. Martinelli, and M. Cernak, “A bin encoding training of a spiking neural network-based voice activity detection,” *arXiv preprint arXiv:1910.12459*, 2019.
- [57] Anonymous, “Gau: Gated attention unit for efficient and scalable sequence modeling,” *arXiv preprint arXiv:2102.11582*, 2021.
- [58] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, “CspNet: A new backbone that can enhance learning capability of cnn,” *arXiv preprint arXiv:1911.11929*, 2019.
- [59] G. Gao, Z. Xu, J. Li, J. Yang, T. Zeng, and G.-J. Qi, “Ctcnet: A cnn-transformer cooperation network for face image super-resolution,” *arXiv preprint arXiv:2204.08696*, 2022.
- [60] MyBib, “Mybib –a new free apa, harvard, & mla citation generator.” <https://www.mybib.com/>, 2024.
- [61] S. Jiang, J. Zhang, J. Feng, L. Zheng, and L. Kong, “Attentive multi-layer perceptron for non-autoregressive generation,” in *Machine Learning and Knowledge Discovery in Databases: Research Track (ECML PKDD 2023)*, pp. 612–629, Springer, 2023.

- 
- [62] K. D. Kihm, “Optical serial sectioning microscopy (ossm),” in *Near-Field Characterization of Micro/Nano-Scaled Fluid Flows*, pp. 29–53, Springer, Berlin, Heidelberg, 2011.
- [63] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, *et al.*, “Opt: Open pre-trained transformer language models,” *arXiv preprint arXiv:2205.01068*, 2022.
- [64] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu, “Roformer: Enhanced transformer with rotary position embedding,” *arXiv preprint arXiv:2104.09864*, 2021.
- [65] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai, “Gqa: Training generalized multi-query transformer models from multi-head checkpoints,” *arXiv preprint arXiv:2305.13245*, 2023.
- [66] N. Shazeer, “Fast transformer decoding: One write-head is all you need,” *arXiv preprint arXiv:1911.02150*, 2019.
- [67] B. Zhang and R. Sennrich, “Root mean square layer normalization,” *arXiv preprint arXiv:1910.07467*, 2019.
- [68] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” *arXiv preprint arXiv:1607.06450*, 2016.
- [69] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021.
- [70] Y. Wang, Y. Sun, W. Song, S. Gao, Y. Huang, Z. Chen, W. Ge, and W. Zhang, “Dpc-net: Dual path multi-excitation collaborative network for facial expression representation learning in videos,” *arXiv preprint arXiv:2312.00752*, 2022.
- [71] H. Wang, B. Li, S. Wu, S. Shen, F. Liu, S. Ding, and A. Zhou, “Rethinking the learning paradigm for dynamic facial expression recognition,” *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [72] Y. Wang, Y. Yang, Z. Li, J. Bai, M. Zhang, X. Li, J. Yu, C. Zhang, G. Huang, and Y. Tong, “Convolution-enhanced evolving attention net,” *journal name*, 2022.
- [73] S. Arlot and M. Lerasle, “Why v=5 is enough in v-fold cross-validation,” *arXiv preprint arXiv:1210.5830*, 2014.

- 
- [74] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” *arXiv preprint arXiv:1512.02325*, 2015. Revised Dec. 2016.
  - [75] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” *arXiv preprint arXiv:1708.02002*, Feb 2018.
  - [76] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” *arXiv preprint arXiv:1804.02767*, Apr 2018.
  - [77] X. Zhou, D. Wang, and P. Krähenbühl, “Objects as points,” *arXiv preprint arXiv:1904.07850*, Apr 2020.
  - [78] M. Tan and Q. V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” *arXiv preprint arXiv:1905.11946*, Sep 2020.
  - [79] Z. Cai and N. Vasconcelos, “Cascade r-cnn: Delving into high quality object detection,” *arXiv preprint arXiv:1712.00726*, 2018.
  - [80] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, “Learning social etiquette: Human trajectory understanding in crowded scenes,” in *European Conference on Computer Vision*, pp. 549–565, Springer, 2016.
  - [81] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” *arXiv preprint arXiv:2005.12872*, 2020.
  - [82] Z. Tian, C. Shen, H. Chen, and T. He, “Fcos: Fully convolutional one-stage object detection,” *arXiv preprint arXiv:1904.01355*, 2019.
  - [83] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, “Makeittalk: Speaker-aware talking-head animation,” *arXiv preprint arXiv:2004.14289*, Feb 2021. Submitted on 27 Apr 2020 (v1), last revised 25 Feb 2021 (this version, v3).
  - [84] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, “Face2face: Real-time face capture and reenactment of rgb videos,” *arXiv preprint arXiv:2007.14875*, Jul 2020. Submitted on 29 Jul 2020.

- 
- [85] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, “Cvt: Introducing convolutions to vision transformers,” *arXiv preprint arXiv:2103.15808*, Mar 2021. Submitted on 29 Mar 2021.
- [86] J. Yu, W. Han, A. Gulati, C.-C. Chiu, B. Li, T. N. Sainath, Y. Wu, and R. Pang, “Dual-mode asr: Unify and improve streaming asr with full-context modeling,” *arXiv preprint arXiv:2010.11439*, Oct 2020. Revised Jan. 2021.
- [87] K. Shiohara and T. Yamasaki, “Detecting deepfakes with self-blended images,” *arXiv preprint arXiv:2304.08747*, Apr 2023.
- [88] A. Haliassos, R. Mira, S. Petridis, and M. Pantic, “Leveraging real talking faces via self-supervision for robust forgery detection,” *arXiv preprint arXiv:2201.07234*, Oct 2022. Submitted on 18 Jan 2022 (v1), last revised 21 Oct 2022 (this version, v3).
- [89] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, “Lips don’t lie: A generalisable and robust approach to face forgery detection,” *arXiv preprint arXiv:2012.05534*, Aug 2021. Submitted on 14 Dec 2020 (v1), last revised 15 Aug 2021 (this version, v3).
- [90] Y. He and Y. Wang, “Rawnet: Fast end-to-end neural vocoder,” *arXiv preprint arXiv:1904.03229*, Mar 2023. Submitted on 10 Apr 2019 (v1), last revised 10 Mar 2023 (this version, v2).
- [91] J. Yamagishi, X. Wang, V. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, “Asvspoof 2021: Accelerating progress in spoofed and deepfake speech detection,” *arXiv preprint arXiv:2109.00537*, Sep 2021.
- [92] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *arXiv preprint arXiv:2110.13900*, Jun 2022.
- [93] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, P. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, “Xls-r: Self-supervised cross-lingual speech representation learning at scale,” *arXiv preprint arXiv:2111.09296*, Dec 2021.

- 
- [94] Y. Liu, B. Schiele, and Q. Sun, “An ensemble of epoch-wise empirical bayes for few-shot learning,” *arXiv preprint arXiv:1904.08479*, Jul 2019.
  - [95] X. Liu, P. He, W. Chen, and J. Gao, “Multi-task deep neural networks for natural language understanding,” *arXiv preprint arXiv:1901.11504*, May 2019.
  - [96] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” *arXiv preprint arXiv:1709.01507*, 2018.